



This is a digital copy of a book that was preserved for generations on library shelves before it was carefully scanned by Google as part of a project to make the world's books discoverable online.

It has survived long enough for the copyright to expire and the book to enter the public domain. A public domain book is one that was never subject to copyright or whose legal copyright term has expired. Whether a book is in the public domain may vary country to country. Public domain books are our gateways to the past, representing a wealth of history, culture and knowledge that's often difficult to discover.

Marks, notations and other marginalia present in the original volume will appear in this file - a reminder of this book's long journey from the publisher to a library and finally to you.

Usage guidelines

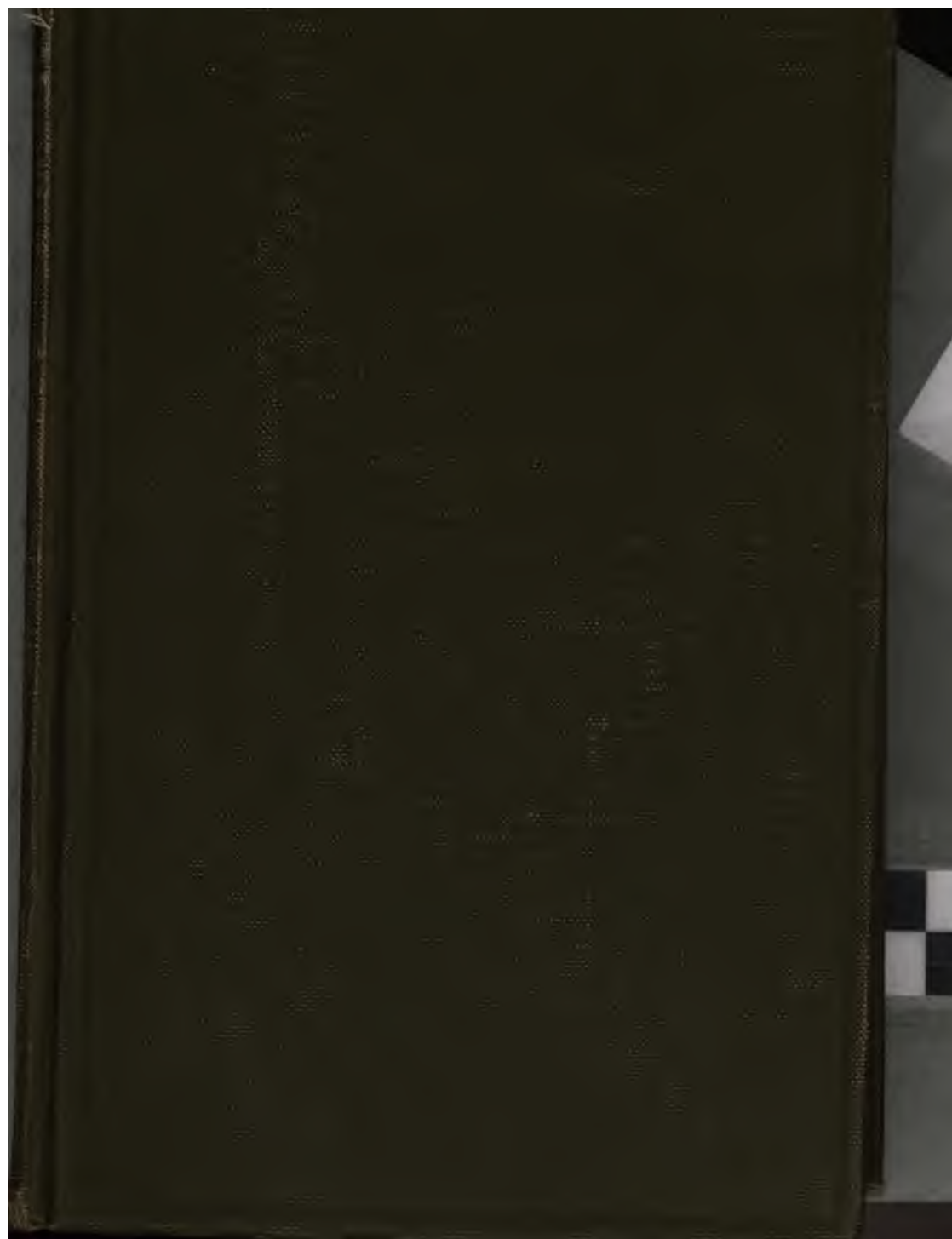
Google is proud to partner with libraries to digitize public domain materials and make them widely accessible. Public domain books belong to the public and we are merely their custodians. Nevertheless, this work is expensive, so in order to keep providing this resource, we have taken steps to prevent abuse by commercial parties, including placing technical restrictions on automated querying.

We also ask that you:

- + *Make non-commercial use of the files* We designed Google Book Search for use by individuals, and we request that you use these files for personal, non-commercial purposes.
- + *Refrain from automated querying* Do not send automated queries of any sort to Google's system: If you are conducting research on machine translation, optical character recognition or other areas where access to a large amount of text is helpful, please contact us. We encourage the use of public domain materials for these purposes and may be able to help.
- + *Maintain attribution* The Google "watermark" you see on each file is essential for informing people about this project and helping them find additional materials through Google Book Search. Please do not remove it.
- + *Keep it legal* Whatever your use, remember that you are responsible for ensuring that what you are doing is legal. Do not assume that just because we believe a book is in the public domain for users in the United States, that the work is also in the public domain for users in other countries. Whether a book is still in copyright varies from country to country, and we can't offer guidance on whether any specific use of any specific book is allowed. Please do not assume that a book's appearance in Google Book Search means it can be used in any manner anywhere in the world. Copyright infringement liability can be quite severe.

About Google Book Search

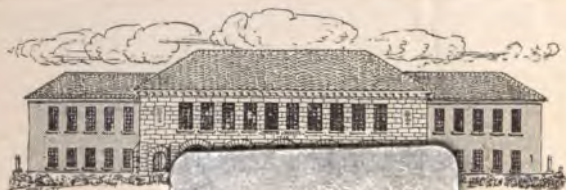
Google's mission is to organize the world's information and to make it universally accessible and useful. Google Book Search helps readers discover the world's books while helping authors and publishers reach new audiences. You can search through the full text of this book on the web at <http://books.google.com/>



371.27
G822



STANFORD UNIVERSITY LIBRARY



SCHOOL

ION

Handwritten text, mostly illegible due to extreme fading and bleed-through from the reverse side of the page. The text appears to be organized into several paragraphs or sections, but the specific words and sentences cannot be discerned.

1730

**FUNDAMENTALS OF
EDUCATIONAL MEASUREMENT**

FUNDAMENTALS OF EDUCATIONAL MEASUREMENT

**WITH THE ELEMENTS OF
STATISTICAL METHOD**

BY

CHESTER ARTHUR GREGORY, Ph.D.

**PROFESSOR OF SCHOOL ADMINISTRATION AND DIRECTOR OF THE BUREAU
OF EDUCATIONAL RESEARCH IN THE UNIVERSITY OF OREGON**



**D. APPLETON AND COMPANY
NEW YORK LONDON**

1922

**COPYRIGHT, 1922, BY
D. APPLETON AND COMPANY**

360989

1959 JUL 1 1907 MAT 2

PRINTED IN THE UNITED STATES OF AMERICA

PREFACE

THIS book is an attempt to bridge, in part, a gap between theory and practice in educational tests and measurements. The leaders in the movement for a quantitative study of educational problems have advanced so far into the theory and more complex practical phases of the work that a very large percentage of the teachers and students have considerable difficulty in following them. Most of the books on the subject have either been more or less technical, pre-supposing considerable training on the part of the readers in the field, or they have been largely manuals of directions for giving the tests and scoring the papers, with little reference to the theory and to the problems which have confronted those attempting to measure educational processes and products. This book deals with the processes and problems in a somewhat evolutionary way so that the teachers and students may see the order in which the problems have arisen and the attempted solutions of them. A mere manual of directions for giving tests and scoring the papers will not develop a professional spirit among teachers in this field. They must understand the fundamental principles, or the work becomes purely mechanical and non-professional. It has been the aim of the author to present the fundamental principles in non-technical language, as far as it is possible to do so, and to confine the statistical treatment of the data almost entirely to simple operations in arithmetic.

The author has drawn freely on the works of Drs. Edward L. Thorndike, Leonard P. Ayres, Harold O. Rugg,

W. C. McCall, Lewis M. Terman, and others to which he desires to give due recognition.

The author desires also to thank Prof. H. R. Douglass and Dr. B. W. DeBusk for reading parts of the manuscript and especially to thank Professor W. C. McInnis of the Eugene High School for the careful reading of the manuscript and the helpful suggestions made.

C. A. G.

CONTENTS

	PAGE
PREFACE	v
 CHAPTER	
I. INTRODUCTION	1
Efficient American Citizens	3
Progress Conditioned by Ability to Measure	4
Working Hypothesis for Acquiring Efficiency	8
II. EFFICIENT THROUGH MEASUREMENTS	10
Statement of and Adherence to Definite Aims	10
Value of General Aims Limited	11
A Danger to Be Avoided	12
The Technical Scientist and Educational Creeds	12
The Elimination of Waste	14
Limited Quantities of Things Make Measure- ments Necessary	14
Waste Not Peculiar to American Schools	14
Human Energy Must Be Conserved	15
Opportunities for Waste in Education	15
The Business Man Eliminates Waste through Accurate Knowledge of His Processes and Products	16
Wastes in Education Many and Varied	17
Four Economies in Education	18
Placing Education on a Factual Basis Through Edu- cational Measurements	20
Two Kinds of Opinions and Their Uses	21
Opinions Worthless in the Face of Facts	22
Inductive Methods Not Used in Pedagogy	24
Lines of Demarcation between Knowable Facts and Philosophical Opinions Must Be Sharply Drawn	24

CHAPTER	PAGE
Data on Simplest School Room Procedure Lack- ing	25
Cultural Development Supplemented by Mechan- ical Phases of Education	26
No Excuse for Mere Opinions Where Facts Are Available	26
Relation of Time Consumed to the Finished Product	28
The Public Is Asking for a Ledger Account of Our Business	31
Inability to Show Facts Has Worked a Hard- ship on the Teaching Profession	32
The Establishment of Definite Standards	34
Three Kinds of Standards Needed	36
Standards as to Quantity	36
Standards as to Time	38
Standards as to Quality	39
Standards Changing	39
Profiting by the Experience of Those in Other Occu- pations and Professions	39
The Manufacturer Has Specifications	41
Meeting Some Objections to Educational Tests and Measurements	43
1. Tests Will Not Endure	44
2. Child Mind too Complicated to Measure	44
3. The Judgment of a Competent Man Better than Scales	45
4. Tests Tend to Reduce All Educational Work to a Dead Level with no Allowance for Indi- vidual Differences	45
5. Tests Measure so Small a Part of Intellectual Life that They Are Not Indicative of Gen- eral Ability	47
Keeping an Accurate Record of All Methods Tried and Progress Made	49
The Cultivation of the Confidence and the Utilization of the Support of the Public	49

CONTENTS

ix

CHAPTER	PAGE
Teachers Must Know Why Sweeping Changes Are Made	49
The Public Is Interested in Education . . .	51
Fields of Educational Tests and Measurements . .	53
III. THE MEASUREMENT OF INTELLIGENCE	56
General Statement of the Problem	56
Why Work in Psychological Measurements Was Re- tarded	60
Effects of Wundtian Laboratories	60
What Is General Intelligence?	63
Is Intelligence a General Faculty of the Mind? . .	65
Inability to Define Intelligence Accurately Does Not Prohibit Measurements	67
The Binet Tests	67
A Tabular Synopsis of the Binet Scale, 1911 Edition	68
The Binet Tests Had Many Innovations	72
The Constituent Functions of Intelligence Must Be Brought into Play	72
The Kind of Mental Functions Brought Into Play	74
Establishment of a Zone of Normality	75
Criteria for Separating the Normal from the Sub- normal	79
Are Differences in Intelligence One of Degree or of Kind?	82
Choosing Tests to Measure Intelligence	83
Tests Must Not Be Influenced by External and Chance Conditions	84
Only Those Tests Must Be Chosen That Afford a Decided and Reliably Symptomatic Value, General Applicability and Possibility of Objective Evolution	85
Tests Must Not Depend Too Much on the Ability to Use Language	85
Determining the Age to Which a Test Should Be Assigned	86
Problems in Scoring	87

CHAPTER	PAGE
The All-or-none Method in Scoring	88
Shall a Child Be Required to Pass All the Tests at Each Age Level?	88
With What Tests Shall the Examination Begin?	89
The "At Age" and the Normal Child	90
Binet Tests Give More Than a Composite Picture	90
The Coefficient of Mental Age and the Intelligence Quotient	90
Limitations of the Tests	91
The Age of Mental Maturity	91
Criticisms of the Binet Tests	93
Limits of Traits Not Determined by the Binet Scale	94
The Binet Scale Criticised on Other Points	95
Some Favorable Criticisms of the Binet Scale	99
Other Problems Confronting Those Testing Intel- ligence	100
Does the Defective Progress Normally to a Cer- tain Point and Then Suffer Arrest or Has Mental Development Been Retarded from Birth?	100
Does the Defective Child Have the Same Mental Equipment as the Normal Child of the Same Mental Age?	101
Do Feeble-Minded Children Mature Mentally at the Same Chronological Age as Normal Chil- dren?	101
Are Subnormal Children Equally Deficient in all Abilities?	101
IV. THE MEASUREMENT OF INTELLIGENCE— <i>Continued</i>	104
The Extension and Revision of the Binet Scale and Other Measures of Intelligence	104
The Stanford Revision of the Binet Scale	104
To Find the Mental Age of a Child by the Stan- ford Revision	106
The Picture Completion Tests	107
The Form-board for Measuring Intelligence	108

CONTENTS

xi

CHAPTER	PAGE
A Scale of Performance Tests	109
A Point Scale for Measuring Mental Ability	111
Proposed Reorganization of the Binet Scale by Meumann	117
Endowment or Intelligence Tests Proper	117
Tests of Development in the Narrower Sense	118
Tests of Environment	118
Wallin's Criticisms	118
Other Tests Devised to Measure Intelligence	119
Group Intelligence Tests	122
Principles Involved in the Selection of Group Tests	123
Requirements of Group Tests Are Many	124
The Terman Group Tests of Mental Ability	125
The National Intelligence Tests	127
The Haggerty Intelligence Examinations	129
The Otis Group Intelligence Scale	131
The Dearborn Group Intelligence Tests	132
Uses of Intelligence Tests	132
Summary and Evaluation of the Measurements of Intelligence	136
The Methods Are Yet Crude	137
<hr/>	
V. THE NEED FOR DEFINITE MEASUREMENTS OF SCHOOL ACHIEVEMENTS	147
Time Consumed in Giving Examinations	147
Attitude of Pupils and Teachers Toward Examinations	148
The Marking System Now in Vogue	150
Scientific Measurement of School Achievements is new	152
Those in the Profession Must Take the Initiative for Improvement	153
Purposes of Educational Tests Are Not Generally Understood	154
The Problem to Be Solved	154
What School Achievement Tests Measure	155

CHAPTER	PAGE
Experimental Evidence to Show that School Marks Are Inadequate	156
Marking System Inefficient Because it Does Not Indi- cate Progressive Degrees of Merit	158
Definition of a Scale	159
A Valid Scale Must Have Equal Steps and Each Step Bear a Definite Relation to the Zero point	161
More Exact Measurements Will Make Education a Science	164
Supervision Improved as Ability to Measure Increases	165
Our Educational Scales Have Been Subjective	167
Tests Do Not Indicate the Cause of Conditions	168
How Standard Tests Differ from Ordinary Examina- tions	169
Illustrating the Law of the Single Variable	172
Time of Day When a Test Should Be Given	174
Number of Times a Test Should Be Given	176
How Standardized Tests are Helpful in Improving Instruction	177
What Kind of School Achievement Tests is Most Important?	178
 VI. THE CLASSIFICATION OF SCHOOL ACHIEVEMENT TESTS AND THE FUNDAMENTAL PRINCIPLES FOR DESIGN- ING THEM	180
Diagnostic vs. General Tests	180
Degree to which Tests are Diagnostic	181
Formal Tests and Reasoning Tests	185
Rate Tests and Development Tests	186
Quality, Difficulty, and Time or Amount Tests	187
Measurements by Opinion, Comparison and by Stand- ard Tests	189
Classification by Educational Tests	190
Educational Age and Mental Age Compared	192
Accomplishment and Educational Quotients	193
Principles for the Choice of Subject-Matter for Edu- cational Tests and Scales	194

CONTENTS

xiii

CHAPTER	PAGE
Information Desired Determines the Type Subject-Matter	196
Some Characteristics of an Ideal Educational Scale	196
The Establishment of a Zero Point	196
Making the Steps of Equal Magnitude	202
The Scale Must Measure the Desired Educational Product	211
The Test Must Be so Simple in Its Application That It Is Adapted to the Classroom	211
Tests Must Not Require an Undue Amount of Time in Administration	212
VII. SCORING THE TESTS AND THE TREATMENT OF THE MEASURES	214
The Problems of Scoring	214
Example in the Development of a Geography Test	214
State Examination Questions Examined	216
The Division Chosen	216
The Selection of Cities to be Used in the Test	217
Advantages of Tests Thus Designed	221
Time is Saved for the Pupil	221
Personal Equation is Eliminated	221
Much Time is Saved in Scoring	221
More Ground May be Covered by a Test Thus Designed	221
Pupils Will Review the Whole Field in Preparing for the Examination	222
The Determination of Scores	222
Where There is Choice Among More Than Two Answers	223
Some Objections to Tests of This Kind	224
Effect of Incorrect Statements Being Placed Before the Student	225
The Values to be Assigned to Scores	226
General Problem of Weighting Scores	227
By the Teacher's Judgment	228

CHAPTER	PAGE
By Weighting the Parts According to the Distribution of Abilities as Shown by the Normal Frequency Curve	228
Accumulation Scores and Scores of Greatest Difficulty	229
VIII. THE MEASUREMENT OF EDUCATIONAL PROCESSES AND PRODUCTS IN FIVE FIELDS OF PUBLIC SCHOOL WORK	232
Measures of the Materials of Instruction	233
Determination of a Spelling Vocabulary	233
A Study of Reading and Spelling Vocabularies of the Books Used in the First Three Grades	237
The Problem	237
Terms Used in the Study	238
The Contents of Three American Histories	240
The Measurements of the Physical Growth of School Children	249
The Measurements of the Money Cost of Education	251
The Measurements of School Buildings	251
The Measurements of Retardation, Acceleration, and Elimination	253
IX. EDUCATIONAL STATISTICS, GENERAL STATEMENT	260
The Use of Statistics in Other Fields	261
The Question of Error	262
Distribution of Measures About a Point of Central Tendency	262
Educational Measurements Compared With Measurements in Other Fields	264
Quantities Measured Indirectly	266
Definition of Statistics	267
Laws of Statistical Regularity	268
Methods of Statistics	269
Limitations of Statistics	270
Standard of Accuracy	270
Compensating vs. Cumulative Errors	270
Discrete and Continuous Series	271
Undistributed Measures	271

CONTENTS

xv

CHAPTER	PAGE
Rules for Tabulating Data	272 -
General Directions for Making a Scale and Curve- Plotting	273
Designating the Class Intervals	274
Analysis of Results	275
The Need for Understanding Statistical Formulas	276
 X. THE MEASUREMENT OF CENTRAL TENDENCY, OR AVER-	
AGES	278
Averages	279
The Arithmetic Mean	279
Computation of the Arithmetic Mean by the Short Method	283
Summary of Steps in the Computation of the Arith- metic Mean by the Short Method	286
The Mode	286
The Median	287
The Spread of the Score Interval Commonly Used in Statistics and School Practice	289
What Formula Shall Be Used in Computing the Median?	290
Computation of the Median, Simple Distribution	291
Case 1. The Number of Items in the Distribution Is Odd	291
Case 2. The Number of Items or Scores Is Even	292
When the Distribution Is Complex	293
Case 3. Where More Than One Pupil Make the Same Score and the Data Are Grouped in Class Intervals	293
Case 4. Where the Median Falls in the 100 or the Zero Class Interval	297
Case 5. When the Partial Sum is the Half Sum and There Is no Correction	298
Case 6. Where the Measures Are Discrete	299
Case 7. Where the Median Falls in the Class Interval Containing no Cases	300

CHAPTER	PAGE
Summary of Steps in the Computation of the Median	301
Comparison of the Arithmetic Mean, Median, and Mode	301
Quartiles and Percentiles	303
XI. MEASUREMENTS OF DISPERSION, OR VARIABILITY	304
How Variability Is Measured	305
The Measures of Absolute Variability	305
Computation of Mean Deviation	308
Computation of Mean Deviation: Data Grouped in a Frequency Distribution	310
Summary of Steps in the Computation of the Mean Deviation by the Short Method	314
The Computation of Standard Deviation	315
The Computation of Standard Deviation by the Short Method	315
Summary of Steps in the Computation of Standard Deviation by the Short Method	316
The Coefficient of Variability	321
XII. THE MEASUREMENT OF RELATIONSHIP, OR CORRELATION	324
Need for Measures of Relationship	324
Illustrating the Computation of the Coefficient of Correlation: Data Simple and Ungrouped	328
Illustrating the Computation of the Coefficient of Correlation: Data Complex and Grouped in Class Intervals	331
Summary of Steps in the Computation of the Coefficient of Correlation	335
Illustrating the Computation of the Coefficient of Correlation by the Short Method (Adapted from Ayres) ¹	337
Correlation Between Two	344
thod	

CONTENTS

xvii

CHAPTER	PAGE
Many Pairs of Values With Data Grouped in Class Intervals	346
Finding the Equation of a Straight Line of Regression	351
Pearson's Equation for a Line of Regression . . .	353
The Reliability of the Correlation Coefficient . .	360
Spearman's Method of Rank Correlation . . .	362
TABLE OF SQUARES AND SQUARE ROOTS	367
INDEX	377

FUNDAMENTALS OF EDUCATIONAL MEASUREMENT

CHAPTER I

INTRODUCTION

The crying need of the hour is for efficiency. The fact that there are few universal standards for efficiency does not dampen the ardor of those demanding it. The commercial world will tolerate almost anything but what it considers inefficiency; the efficiency, or inefficiency, of our railway systems constitute no small part of the conversation of the shippers and the traveling public. In politics we talk about an efficient administration or an efficient public servant; in education, the public wants an efficient school system.

The subject of efficiency has occupied the center of the stage at every noted gathering of educators in the United States for the last decade. A belief is current that the education of the day is far short of its optimum efficiency; that there is something wrong with the public schools. Because of this criticism the elementary schools have been compelled to face, not, as of old, a criticism as to their rights to existence or their right to receive public support, but an adverse criticism as to their efficiency as measured by their human product.

The educator can withstand almost any kind of criticism but that which decries the efficiency of his school. Effi-

ciency is one of the chief goals for which he is striving. It is the balance in which his success or failure must be weighed. Books, consisting of hundreds of pages each, have, as their chief contents, data dealing with the efficient administration of the various phases of school systems.

Recently there has come from the press a five-hundred page report on the Gary public schools made by the General Education Board. The makers of this survey say in the opening pages of their report:¹ "The public is interested in knowing whether the Gary schools are efficient or inefficient as now conducted. The public is also interested in knowing whether such a plan is sound or unsound. The present study tries to do justice to both points."

The National Society for the Study of Education considered efficiency of such importance that it devoted Part II of the *Fourteenth Yearbook* to the "Methods of Measuring Teachers' Efficiency." Scores of other books on education make efficiency no small part of their content.

The efficiency programme is not confined to the public schools. Education is but one phase of the newer social economy. When one runs through the card index of the library or consults the *Reader's Guide*, he is surprised to see the number of books and magazine articles dealing with the subject of efficiency in all lines of work. The states of the Union appoint commissions on efficiency and economy. Economy and efficiency in government service becomes a part of the report of the President of the United States to Congress. Books on efficiency are written under such titles as the following: "Modern Methods in the Office, How to Cut Corners and Save Money"; "Economics of Efficiency"; "Psychology for Business Efficiency"; "Efficiency as a Basis for Operation and Wages"; "Motion

¹ Stuart A. Courtis, *The Gary Public Schools, Measurement of Classroom Products* (General Education Board, 61 Broadway, New York City), Introduction, p. 8.

Study, A Method for Increasing the Efficiency of the Workman"; "Fatigue and Efficiency"; "The Price of Inefficiency"; "The Human Machine and Industrial Efficiency"; "A Symposium on Scientific Management and Efficiency in College Administration"; and hundreds of other books and articles dealing with this subject.

The spirit of the new education is that of social efficiency. Subjects will no longer find an excuse for being in the course of study in the words of the old song, "We're here because we're here," but each subject must prove that it is one of the best things that can be taught in the few short years of a child's school life.

The mechanic has a perfectly definite way of measuring efficiency. *It is the ratio that the useful work done by a machine bears to the total work done on it.* No machine is one hundred per cent efficient, else a perpetual-motion machine would be possible. There is always a waste. The most efficient machine is the one which does the maximum amount of work with a minimum of waste.

The efficiency of a school system is likewise a ratio. It is the ratio that the time, energy, and money spent bear to the finished product. The finished product in the case of the machine, that is, the work done, is measured in some convenient unit of work, as the foot-pound, the foot-poundal, the erg, or the kilogrammeter. Unfortunately, the products of a school system are measurable in no such definite terms. Descriptive adjectives with vague meanings usually constitute the best and only measuring sticks we have.

Efficient American Citizens.—It is said that one of the chief businesses of the American schools is to prepare individuals for efficient membership in American society. There are at least three qualifications for such membership: (1) An ability to execute effectively the formal and informal duties of citizenship and carry the burdens of

political responsibility; (2) an ability to labor and produce so effectively that one is able to carry his own economic load; (3) an ability to utilize one's leisure time and to function in an individual capacity without infringing on the rights of others or of society at large.²

Few would disagree with these qualifications. It is only when we attempt to particularize and say just *what* and *how much* an individual must have to execute the formal and informal duties of citizenship, carry the economic load and perform the other duties mentioned above, that differences of opinion arise.

All agree that the schools should turn out boys and girls efficient in arithmetic, language, civics, good character, and all the rest, but there is little agreement as to just what one must do to become efficient in these things. Standards of efficiency and methods of reaching them have not always been based on scientific facts which might be verified in the laboratory. Before one can have an intelligent conception of efficiency, there must be some way to evaluate, to measure it. In fact, one cannot conceive of efficiency except in terms of quantity.

It is the purpose of this treatise to set forth in a non-technical way, as far as possible, some of the benefits to be derived and the difficulties to be overcome in the use of measurements in the field of education.

Progress Conditioned by Ability to Measure.—Progress in civilization has depended very largely on our ability to measure. James Watt, for instance, could not make a steam engine until men were able to make measurements so exact that a cylinder and a piston could be built that were steam-tight and yet allowed free play. The automobile of to-day had to wait until men could measure the five-thousandth part of an inch, and the ship's chronom-

² Alexander Inglis, *Principles of Secondary Education*, p. 342.

eter until they could measure distance five times more minute than that.

New measures are constantly coming into use. They are no longer restricted to length, area, weight, and volume. New commodities such as electric currents, light, heat, and refrigeration are measured and sold on the markets to-day just as other commodities are sold. Progress in these lines has been conditioned by our ability to devise new measuring sticks to measure them.

The physical sciences have taken every precaution to standardize their units of measurement. At the United States Bureau of Standards in Washington one finds instruments that will weigh with accuracy down to $\frac{1}{50,000,000}$ of an ounce; high-temperature thermometers that will register accurately 1,000 degrees above zero Fahrenheit and pentane thermometers registering 300 degrees below zero Fahrenheit; saccharimeters measuring the impurities in sugar by the twist of light waves which are allowed to pass through a sugar solution; Emery testing machines having a compressing power of 2,300,000 pounds and a pulling power of 1,150,000 pounds per square inch.²

It is no longer a matter of opinion as to the strength of a steel girder, for instance, or the resistive power of a steel rail. The scientist now speaks with authority along these lines. Natural science has made its gains by substituting facts for opinions, and units of measure for mere guess-work. Herschel says: "Numerical precision is the soul of science."

When the teaching profession enters the stage where its data and conclusions can be presented in quantitative as well as qualitative terms, it has entered upon a most important stage of development.

Men in all businesses and professions are becoming quan-

²"A Wonderland of Science," *National Geographic Magazine*, Vol. 27, pp. 153-169.

titative thinkers. Education is not an exception. Educators are seeking to verify and, in some cases, to refute the established beliefs concerning the effects of educational forces upon human nature. Dogmatic and authoritative control of education is going the way of all mere authority and dogma in human affairs. The "popular guessing contests" that have been going on in education as to which processes are the best, and what products are obtained from them, are giving way to experimentally determined facts. It means that education is emerging from among the vocations and taking its place among the professions.

In the evolution of our civilization, one form of human activity after another has been subjected to exact measurement and made to yield its quota of natural law. This movement in education is but a part of a larger one which in recent years has extended the scope of the applied sciences and utilized scientific principles in the improvement of many lines of human endeavor.

The American people are an extremely practical folk. Contemporaneous American life manifests an abiding faith in practicality and efficiency. The American people believe in a creed or philosophy of life that really "works" when applied to practical situations. They are pragmatists. They are looking away from first things, principles, categories, supposed necessities, and looking towards last things, fruits, consequences, facts.⁴ We have talked much of the aims of education. Now we are asking for results. We are going to judge the efficiency of a school system in terms of the results and not in terms of its aims.

For many years we have measured, in a way, the efficiency of higher education by the definition of what constitutes an institution of higher learning, which we find in the laws of the various states, in the definitions of the

⁴ William James, *Pragmatism*, p. 54.

U. S. Bureau of Education, the Carnegie Foundation for the Advancement of Teaching, and many voluntary associations that deal with this phase of education. These definitions cover such topics as: entrance requirements, endowments, income, curricula, the school plant, time allotments, and the qualifications of the teaching staff. Such standards measure the equipment and materials for instruction but not the product.

Much emphasis has been laid on the statement that the teacher's power is exerted in one generation, while that of his students is exerted in another; that for him who teaches there is no final measure of the day's work; that it is beyond his vision in time and place. The next generation may attempt a full estimate of his labor, but he himself cannot.

Teachers have accepted these general statements more literally than the facts warrant. It is true that the *full* power of one mind over another cannot be measured. Nevertheless, if the intricate ministry of teaching is to become anything more than a crude art, where blind faith, subtle intuition, and crude methods of trial and error, are the rules of procedure, standards must be set up and measurements must be made, not in the next generation, but in the midst of the educational processes now going on.

The ideal condition would be for the teacher to see the man of his moulding walking about full-grown among his neighbors and performing his daily duties and graces. No other measure of one's work equals the sight of the product put to its full uses. It is the best corrective of one's blunders, the quickest encouragement to efficient action. Unfortunately, this satisfaction is reserved for the lesser craftsmen of life. We shall have to be content with conditions less favorable. This does not mean, however, that because we cannot measure our educational products completely, we cannot measure them at all. Some phases of

our educational processes and products are quite amenable to measurements. The hazard is too great to wait for the next generation to evaluate our products. Definite, attainable goals, therefore, are being set up. Teachers are being required to find the situation, or series of situations, that will produce the desired results. When a child is put through the education situation, we want to know immediately the quality and amount of change made. If the results are not satisfactory, we put him through the situation again or make a new situation that will produce the change desired.

Working Hypotheses for Acquiring Efficiency.—No attempt will be made here to define an efficient school system. Enough has been said about efficiency to give the general direction of the goal toward which we are striving. No attempt will be made to discuss educational creeds unless measurements become a factor in determining them. The chief problem, rather, is this; having decided upon a desired result, how may measurements be utilized to bring about that result with the minimum expenditure of time, money, and energy? Our hazy ideas about an efficient school system will be clarified only by approaching a little closer to the goal sought. While we do not know exactly what an efficient school system is, we know the road that leads to it.

Whatever the definition of an efficient school system may be, the following propositions may be readily accepted as working hypotheses:

1. *Of two school systems attempting to make efficient American citizens, that system is the better which produces the desired finished product with the minimum waste in time, energy, and money expended.*

2. *More accurate measures than we now have of those things that are measurable will reduce the waste in education and make our schools more efficient.*

It is with the second proposition that we are primarily concerned. An attempt will be made to show how the efficiency of the schools may be increased by a more accurate measurement of certain materials, processes, and products. Three activities are necessary to do this. It is necessary: (1) to persuade the teachers that measurements are beneficial; (2) to make the measurements; (3) to use the results of the measurements as a means for doing better.

Knowledge is the first attribute of the man of science. It is knowledge, born of the travail of thought and experience, that differentiates the physician from the quack; the lawyer from the shyster; the statesman from the demagogue; it is likewise the first indispensable element of educational sanity and progress. Education has hitherto rested on the foundation of custom. It must hereafter rest on a basis of scientific knowledge. Facts alone will enable the educator to keep his balance in the midst of educational upheavals. Because of the lack of scientific information, many theories not justified by systematic observation have been current. As a result, much of the time and energy of teachers have been spent to a great disadvantage; confusion has been produced; and the teaching profession has, at times, been greatly retarded.

In spite of the fact that the educational literature is replete with data and arguments why teachers should use measurements to eliminate waste and do more effective teaching, nevertheless the majority of teachers are not availing themselves of this opportunity. Tradition and habit are still too strong to allow any radical changes. Of those teachers who have used these newly acquired tools, many have treated them as "educational curiosities" rather than as a means for more efficient work. It is because of these conditions that considerable space is given in the next chapter to reasons for the need of measurements.

CHAPTER II

EFFICIENCY THROUGH MEASUREMENTS

In this chapter we shall discuss some reasons why measurements should be made, hoping thereby to persuade a greater number of teachers that measurements are necessary for the most effective teaching. The subject will be discussed under the following headings, all of which have to do with measurements either directly or indirectly. The general thesis is that the schools may be made more efficient by: (1) statement of and adherence to definite aims; (2) the elimination of waste; (3) placing education on a fact basis through educational measurements; (4) the establishment of definite standards; (5) profiting by the experience of those in other occupations and professions; (6) meeting some objections to educational tests and measurements; (7) keeping an accurate record of all methods tried and progress made; (8) the cultivation of the confidence and the utilization of the support of the public.

I. STATEMENT OF AND ADHERENCE TO DEFINITE AIMS

Professor Bobbitt has tersely stated the situation in regard to our aims in education. He says:¹ "We have aimed at a vague culture, an ill-defined discipline, a nebulous, harmonious development of the individual, an indefinite moral character-building, an unparticularized social efficiency, or, often enough, nothing more than escape from a life of work."

¹ John Franklin Bobbitt, *The Curriculum* (Houghton Mifflin Co.), p. 41.

It is evident that, as long as we deal with such rainbow generalities, we can never hope to get definite results. Our results never will be more definite than our aims. One of the first things that the establishment of standard norms will do for the science of education will be to make definite and specific the aims of teaching.

We have a bountiful supply of general aims of more or less value. Education is a training for citizenship; it is to prepare for life; it is to develop power; it is to make cultured men and women; it is to prepare for the vocations; it is to develop a moral individual, endowed with the power of independent thought, with the ability to earn an honest livelihood, with culture, refinement, and a broad and intelligent interest in human affairs. The aims of education are sometimes stated in terms of culture and sometimes in terms of productive efficiency. The tendency now most in favor among progressive school men is that emphasized by the very principle for which social economy stands; that education is to be tested neither by culture in the abstract, nor utility in the concrete, but by the extent to which any slight knowledge, any manual dexterity, and any useful "tricks" of spelling, counting, reading, etc., are assimilated into the organic complex of personal character.

Value of General Aims Limited.—It is easy to agree with general aims but they are of limited value in the school room. Where one is confronted with a particular boy, with a particular task, as adding a ten-figure column, such aims as preparation for life, citizenship, culture, power, etc., seem remote and ineffectual. The controlling purpose of education must be sufficiently particularized that we may know when any part or unit of the aim has been accomplished. An objective such as "good citizenship," for instance, must first be reduced to smaller units such as being a good neighbor, a good parent, a wise and

Does it?

When these be measured — will give of a character

conscientious voter, etc. These units must, in turn, become material for further analysis until we finally get down to definite "working units." The first thing we have to settle is what we want our schools to produce. Our courses of study at the present time easily could be filled with many objective statements of aim which would be so specific that there would be no question whether or not they were carried out in practice.

When we are challenged to justify the increasing cost of education and the multiplication of courses necessary to meet the demands of complex life, we apply the pragmatic test. For instance, what practical difference would it make if this were incorporated into the course, and if that were left out? The progressive teacher is constantly confronted with the decision of accepting an educational fad for a scientific fact. Measurements offer a means by which she can test out the relative values of aims by tracing them to their consequences. Tests and measurements, therefore, become an important guide in evaluating the proposals of educational changes.

A Danger to Be Avoided.—There are many dangers incident to the use of measurements which will be discussed later in this chapter. Only one will be mentioned here, namely, the danger of making the real aims and ends of education subservient to measurements, instead of using measurements to get the desired results. The good physician is interested primarily, not in the medicine, nor in the methods of administering it, but in the effect it will have on the patient, and we should regard him as a poor physician who thinks more of the medicine and the way he gives it than he does of the patient's progress toward recovery.

The Technical Scientist and Educational Creeds.—It is not the business of the technical scientist to tell us what shall be the aims of education. That is a matter of creed.

Those belonging to a creed usually set up an ideal toward which they strive. For instance, there are those in education who look primarily to the subjective results; the enriched mind; refined sensibilities, culture, discipline, and the like. Those belonging to this creed say that education is the *ability to live* rather than the *practical ability to produce*.

Another creed has as its aim *efficient practical action* in a practical world. The individual is educated who can perform efficiently the labors of his calling; who can co-operate effectively with his fellow in social civic affairs. This creed would have science studies in order that the facts may be put to work by the mechanic in his shop, and the farmer on his farm. It would make a survey of the science needs of the community and teach only those things which make a direct contribution to these needs.² The scientist in education has practically nothing to do with the formation of these creeds. He simply says that you must make use of this means, if you wish to reach this or that particular end. But no technical science can decide within its limits whether the end itself is really a desirable one. The technical specialist knows how he ought to build a bridge, or how he ought to dig a tunnel, presupposing that the bridge and the tunnel are desirable. Whether they are desirable or not is a question that does not concern him. He simply says that if you wish this end, then you must proceed this way.

When aims are definite, measurements are made on the materials of instruction so as to meet the specifications set forth in the aim. Everything that does not contribute to the aim is then discarded as material that is superfluous.

² *Ibid.*, pp. 3-4.

II. THE ELIMINATION OF WASTE

Limited Quantities of Things Make Measurements Necessary.—If everything with which human activity is in any way concerned were unlimited, there would be no need for measurements. Even if things were not actually unlimited, if there was always enough of any one of them to be had with little or no expenditure of energy, we would concern ourselves but little about quantity, and hence measurements would be practically unnecessary. The very fact that most of the things with which we deal are limited, forces us to measure in order to best conserve them. A cash register is used because money is limited and therefore must be measured in order to conserve it. Nature has limited the time allotted to man to do things, and, because of this fact, he has built labor-saving machines with which distances are annihilated and work is done as rapidly as possible so that time may be saved. Because our available supply of energy is limited, we must conserve each unit to be utilized in doing the world's work.

One of the fundamental principles to be kept in mind in the solution of educational problems is that *everything with which the educator works—time, energy, money, apparatus, resources of all kinds—are so limited that trusting them in the hands of the ignorant cannot be endorsed, and prodigality with them is crime.*

Waste Not Peculiar to American Schools.—Americans are the most wasteful people in the world. Our bountiful natural resources, our great expanse of land, our seemingly exhaustless supply of those material things that supply human wants have made us wasteful. Suppose, by our unscientific methods in farming, we *did* deplete the soil, was not there plenty of new land awaiting us? What did it matter if millions of dollars' worth of natural gas

was allowed to go to waste in obtaining oil so long as the oil fields yielded a profit? Thousands of tons of fish have been caught and wasted because men have failed to see that the demand might some day overtake the supply. Timber in our great forests has been ruthlessly destroyed because a short-sighted public did not stop to consider that the end was already in sight. Now that the new lands are about all gone, many of the natural resources wasted, the population increasing, and competition growing sharper every day, we have at last begun to realize that we must conserve while there is yet an opportunity.

Human Energy Must Be Conserved.—Because human energy is limited, it becomes necessary to economize and distribute it in such ways that it will accomplish the most. If we put forth more energy than is necessary to do a thing, there is waste. Likewise there is waste if less energy is put forth than is necessary to accomplish the task. We do our most difficult tasks with the least waste of power and energy when we accurately adjust our energies to the task to be done. The ends to be realized many times are remote and complex, and if we use adequate means, distances in space, remoteness in time, units of energy necessary to do the task, quantity of some kind must be taken into account, and this means that measurements must be made. Waste no doubt has occurred in education more because of a lack of scientific methods of conservation than because of any indifference or negligence on the part of teachers.

Opportunities for Waste in Education.—When fully considered, much of the great waste in education is due to our lack of adequate means for placing reliable estimates on our results and processes. We lack in the matter of definite, desirable, attainable goals to be sought through a given topic, or process, or stage of work in a given subject. We have been forced to work in a more or less

blind, do-and-trust-to-luck way. Wherever the application of scientific measurements to the achievements of school children has been made, it has shown that great waste and unbusinesslike methods are being practiced. A school system should meet the same requirements that a business corporation must meet. The output must be commensurate with the expenditure.

Just as in the business world the greatest economies are effected through small savings, so the school must expect to make its greatest gains by checking up the small leaks in time, energy, and expense. A saving of thirty-five minutes a day, for instance, would save a child one school year in eight, or a saving of three and one-half minutes per day would mean a saving of one school month in the course.

The Business Man Eliminates Waste through Accurate Knowledge of His Processes and Products.—The keynote of successful business to-day is accurate knowledge of details applied in such a manner as to eliminate needless waste. The margin between profit and loss is determined by the skill of the manager in effecting small savings. Science applied to the meat-packing industry, for instance, showed that a very big profit could be made by utilizing what had formerly been considered useless. In the business world nothing is left to chance. No important action is based upon vague opinion or untested theory. Exact knowledge must first be obtained. Seven out of ten business failures are due to a lack of definite, obtainable, business knowledge.

A few months ago the writer had an interview with an employee of one of the life insurance companies doing business in the Pacific Northwest. He is what is known as a general agent whose business it is to go about the country and obtain local agents to sell life insurance. By a system of tests, measurements, and personal interviews he is able so to choose his men that eight out of every ten

chosen become successful insurance agents. He said that if he could become 90 per cent efficient, that is, if he were able so to choose his men that nine out of every ten chosen would be successful agents, he would save his company thousands of dollars each year.

Nowhere may the struggle for efficiency be seen more concretely than in industry, for no field has a more constant and compelling motive-profit. Here are exhibited new processes, new labor-saving devices, new methods of planning, more detailed instructions, more exacting records. Astonishing statistics show that the products are doubled and tripled as a result of those methods. There is a decrease in cost for the producer and an increased product for the worker.

Waste of material, waste in effort, in energy, in lives, and in property, is reduced to the minimum in order that the profit may be large.

Wastes in Education Many and Varied.—Because of the many tasks devolving upon the public schools, they have become the foremost instrument of social economy. This added responsibility has increased the chances for waste and inefficiency.

The startling revelation, made a few years ago, that our system of free education was failing to give even complete elementary schooling to the majority of children evoked imperious demands for more real facts. When thousands of children are reaching maturity, unskilled and unwanted, and are pointing an accusing finger at the school they were so glad to leave, the public begins to question the returns for the great sums so lavishly expended on the educational institutions. It begins a careful examination of the waste products, the juvenile delinquent, the youthful criminal, the wayward girl, and the unskilled youth, all of whom are unwanted and ill-adapted for employment. The school must measure quantitatively its

fundamental problems. It must account at every stage for its raw material, its waste products, and its marketable commodities.

The purpose of the school is not that the child shall learn, for he will learn without the school. Learning is a spontaneous process which no lack of schooling can stop and no extent of schooling can do more than modify. Its purpose is to furnish conditions under which the child, through systematic and economic effort, will accomplish more for himself, and that which is accomplished will be of a better quality, and the product will be obtained by him in shorter time and with less expenditure of energy than if he learned it under other conditions.

The principle of economy when applied to the individual child presents many problems of school organization, one of which is that the organization must be such that each child shall have an opportunity for being at his best all the time in all the subjects. The school has but one purpose, the education of children. Consequently, in a strict sense, economic management in education can be defined only as a system of management directed toward the elimination of waste in teaching so that children attending school may be duly rewarded for the expenditure of their time and effort. The point at issue here involves the discovery of processes that, other things being equal, will perform a given task in the smallest amount of time with the minimum expenditure of energy.

Four Economies in Education.—We seek economy in education from four different points of view: (1) economy through the quality of the product; (2) economy in the quantity of the product; (3) economy in the time; and (4) economy in the expenditure of energy. Waste will not be prevented and the maximum accomplished unless these four economies move together, each in harmony with the other, and each, therefore, serving as a check on the

other. Nothing may be gained by "robbing Peter to pay Paul." One fundamental defect in pedagogical thinking has been the overemphasis given to some one feature of human development at the expense of other features no less important. All interests must be carefully considered before we can claim efficiency. In penmanship, for instance, quality demands that the pupil be taught to write *well*, but quantity, time, and the expenditure of energy demand that he do not write *too* well. Quantity and the writer's time demand that the pupil develop speed in writing, but the time and expenditure of energy of the reader are considerations that insist that he do not write *too* rapidly.

In the matter of style, quality demands that the style be most legible. This demand was the cause a few years ago of our substituting the vertical system for the Spencerian. But vertical writing, though most easily read, failed to satisfy the demands of quantity and speed. The attempt to obtain the proper evaluation of these four economies gives rise to experimental problems involving measurements. The efficiency of instruction will be determined, not by developing one phase of the work at the expense of the others, but by a development such that the sum of the net gains in the four economies shall be the maximum.

Education should become efficient as industry is efficient. This means that educators should know what the finished product is. They should be able to gauge the time, quantity, and value of the elements that entered into it. They should see that they actually produce what they say they are going to produce. They must determine the proper ratio of product and time; they must define standards and eliminate waste; if they are to be efficient in conformity to the world-wide ideal of efficiency, they must employ efficiency tools, one of the chief of which is measure-

ment. Education cannot assume the efficiency ideal without adopting its concomitant—measurement.

III. PLACING EDUCATION ON A FACTUAL BASIS THROUGH EDUCATIONAL MEASUREMENTS

If schools are inefficient for any one reason more than another, they are inefficient because of the ignorance of facts concerning their processes and products. Although the problems concerning elementary education have confronted the world for centuries and many educators have attempted their solution, they are still involved in uncertainty and indefiniteness. The statements made on simple practical questions, even among our leading educators, are conflicting to the point of absurdity. Educators are divided into creeds. Those belonging to different creeds are seldom in agreement even on the simplest processes of educational procedure. Of course, there are some phases of education that are matters of creed. What constitutes good citizenship is a matter of creed, but it should be a scientific fact whether the rules of spelling are helpful, or what constitutes reasonable speed and accuracy in adding a column of figures in the fifth grade, or what constitutes a legible specimen of handwriting. Where everything is "guesswork" and there is no proof to offer as to who is right and who is wrong, it is little wonder that the ship of pedagogy is "waterlogged in the sea of opinions." Hume's description of the metaphysical sciences of his day may not inaptly be applied to the condition of current education. He said:³

Even the rabble without doors may judge from the noise and clamour which they hear that all goes not well within. There

³ Robert R. Rusk, *Introduction to Experimental Education* (Longmans, Green & Co.), pp. 1-2.

is nothing which is not the subject of debate, and in which men of learning are not of contrary opinions. The most trivial question escapes not our controversy and in the most momentous we are not able to give any certain decisions. Disputes are multiplied as if everything were uncertain; and these disputes are managed with the greatest warmth, as if everything were certain. Amidst all this bustle 'tis not reason which carries the prize, but eloquence; and no man need ever despair of gaining proselytes to the most extravagant hypotheses, who has art enough to represent it in any favorable colours. The victory is not gained by the men-at-arms, but by the trumpeters, drummers and musicians of the army.

There is much speculation. "I think," "I guess," "It is my opinion" are the characteristic phrases in education. "I know" is a phrase that has scarcely been admitted.

Two Kinds of Opinions and Their Uses.—Generally speaking, we may classify opinions under two general headings: (1) expert opinion, and (2) the opinion of the layman. By expert opinion we mean the opinion of one who, by his long experience and study, is able to base his judgment on data usually not available to, or at least not possessed by, the average individual. For example, J. E. Wooters wanted to determine the dates and events that might be memorized most profitably by students in American history in the seventh and eighth grades. He sent questionnaires to the members of the American Historical Association enclosing a list of 52 dates and requesting that the 20 most important dates in this list be arranged or "ranked" in the order of their importance. If other dates, not given in the list, were, in the judgment of those making the reply, more important than those submitted, they were to be inserted. When the answers were compiled, the date 1776 ranked first in importance, 1492, second, and so on. These men presumably were giving expert opinion. Being historians by profession, it was

presumed that their opinions were the conclusions drawn from the best historic evidence available.⁴

The opinions of both the experts and the laymen may be individual or they may be the combined, average, or median opinions of a group. For example, the judgment of three or four surgeons called into consultation as to whether a certain operation should be performed illustrates group opinion. Generally speaking, group opinion is more reliable than individual, whether it be among experts or laymen. It is less liable to be affected by personal bias and superficial characteristics.

By lay opinion we mean the opinion or judgment of the average individual who expresses himself on the various current problems and topics, scientific and otherwise, with little or no data upon which to base his judgment. This type of evidence has been ruled long since out of court but *not* out of education.

Although the opinion of the layman is of no value for scientific purposes, it is, nevertheless, a formidable factor in education, simply because the margin between technical information and lay opinion is so narrow that the educational expert is not always able to convince the public that his position is right.

Opinions Worthless in the Face of Facts.—In the absence of facts opinions reign supreme. In educational procedure mere opinion has had an all-too-important place on the programme. Where facts are available there is no longer any justification for mere opinion. The following illustration will show the fallacy of basing any procedure on mere opinion when the facts are available. Suppose a group of 40 individuals were asked their opinions of the length of a certain room. Let us suppose that each

⁴ W. C. Bagley, "The Determination of Minimum Essentials in Elementary Geography and History," National Society for the Study of Education, *Fourteenth Yearbook*, Part I, pp. 139-140.

of them is an expert in judging lengths of rooms. Their individual judgments may range all the way from 45 to 50 feet. Then suppose we take the average judgment of the entire group and we find it to be 48 feet. In the absence of a measuring stick this may be considered the best measurement that it is possible to get. Yet it is quite evident that this judgment may not correspond to the facts at all. It is possible that a measuring stick might show the length of the room to be 45 feet 8 inches. It is also quite possible that no one of the experts judged the length to be what the yard stick showed.

Conditions analogous to this have prevailed in education from time immemorial. Either because we did not have the facts or did not care to go to the trouble to get them, false judgments have been allowed to stand. Education and theology are two fields where opinions have reigned supreme. The layman has been free to challenge the judgment of the minister and the teacher because neither could prove, nor disprove the points at issue. The facts upon which judgment should have been made have not been available. Fortunately, however, assertions are no longer the style in educational circles. The scientific spirit of the age demands that assertions be backed up with statistics based upon the results of experiments. *Iipse dixit* will no longer suffice.

Those working to develop measurements realize that only by getting the facts can education become an art and a science, and its practice changed from a vocation to a profession. Just as measurements and the possession of facts developed astronomy from astrology, chemistry from alchemy, and physics from mystery, so they will rescue education from the sea of opinions and place her in the family of sciences where she rightly belongs.

The reason why illusion and unverified hypotheses have so run riot through educational discussion has been the

accepted standards of measurements. We have guessed that we were not making progress in arithmetic. We decided to devote a week to make the work more efficient. We also guessed that pupils were making considerable progress from grade to grade, but we never knew actually what was that progress. We have guessed all along the line. Educators have tried to solve the problem on means of hypotheses based on psychology, whereas facts alone will tell the tale. As a consequence, we have a great mass of philosophical opinions as to what should be done.

Inductive Methods Not Used in Pedagogy.—Pedagogy presents a remarkably anomalous condition, for, as the department that charts the way to the development of the sciences it itself has failed to adopt what it has long been recommending to other scientific pursuits, namely, the inductive method of study. Its work has consisted of opinions based on opinions and, therefore, of a mass of contradictory material. No really sustained forward movement can be expected until conflicting views are subjected to analysis in the light of clear and unmistakable facts. In the expectation that we are able to obtain the genuine and exact knowledge, education will move forward among its sciences. Unless the building of educational opinion is established on verifiable facts which any technically informed person may apply, educational procedure will never become scientific. We must have technical information, the validity of which is indisputable. When this is done, we shall displace imaginative guesswork, and authority be criteria for reshaping the educational policies.

Lines of Demarcation between Knowable Facts and Philosophical Opinions Must Be Sharply Drawn.—It must not be inferred that all problems in education are amenable to measurement. As was indicated above, elementary education presents two distinct types of problems: one

is involved in subtleties and belongs to the department of philosophy; the other is more superficial and is in a large part a question of science. The first includes those factors which relate to the development of character and the cultural phases of one's life. Culture emphasizes the things of the mind and the higher life. It seeks to beget the ability to enjoy the beautiful and the good wherever these may be found. It seeks the ability to think the best thoughts of the best men that we find enshrined in literature. If one would make life worth living, he must partake very largely of cultural things. These things are not directly amenable to measurements, at least not at the present time. On the other hand, the mechanical skills, positive knowledge, most of the things taught in the formal subjects, such as arithmetic, grammar, reading, spelling, etc., readily yield to quantitative treatment. It lies within our reach to ascertain the time consumed by different teachers in obtaining certain positive results and to discover what processes have proved most economical.

Data on Simplest School Room Problems Lacking.—Anyone accustomed to the problems that come up for solution in the course of a day's teaching can think of dozens of questions for which he would like a definite answer, as, for instance: What words should be taught in spelling? How many? What methods should be used? How rapidly should a boy in the fourth grade read, that is, how many words per minute? What is meant by legible handwriting? What dates should be taught in history? Are rules beneficial in the teaching of spelling? Standard tests and scales offer the same effective instruments of research in dealing with these problems as the meter and the gram offer to the student of physics. It is difficult to see how anything can be done for the science of education without them. At present we are absolutely unable to form an intelligent judgment of how much time should

be required to teach any of the simple, formal things in education. Educators are not yet agreed what words should be taught, when to teach them, and how long it should take to do it. It is in these fields that education is inefficient.

Consider - Tolson

Cultural Development Supplemented by Mechanical Phases of Education.—While everyone engaged in measuring recognizes that the highest product of the school training is the development of ideals and character in children, some are becoming convinced that successful work in character-building is absolutely dependent upon the successful work of developing the fundamental skills and the formal phases of education. The studies that have been made show that culture and inspirational work are conditioned by the proper equipment of the child with the mechanical tools by which all mental work is done. It is a mistake, therefore, to think that the time devoted to the mechanical skills is out of harmony with, and antagonistic to, cultural education. On the other hand, it is equipping the child with the tools for appreciating the beautiful and the good.

No Excuse for Mere Opinion Where Facts Are Available.—The time has come in education when mere opinion will satisfy neither the modern school administrator, nor the public. Things in the business world are measured and standardized with a precision not yet reached in education. Even those things which a few years ago were considered incapable of measurement are now measured and standardized so that the public knows exactly what is meant when a thing is spoken of as belonging to a certain class or grade. For instance, oranges and apples are standardized according to the variety and size. When a grocer orders grapefruit and says to the salesman, "Send me a box of 'sixty-fours,'" he knows just what to expect and knows whether the order has been filled prop-

erly when it arrives. In the case of shoes the average individual may not be able always to tell the difference between "firsts" and "seconds" but the expert can, and it isn't mere guesswork on his part. He has perfectly definite standards by which to judge.

If we ask a carpenter to build a room we give him a set of specifications of the length, width, and shape of the room that we desire to have built. When he has finished the job, we can take the specifications and check the finished product to see if it is done according to them.

Any business man who is managing a successful business in which he expects to make profits will have standards from which he will work. He will have measuring sticks to measure the efficiency of his output; he will have units of accomplishment; he will have cost-accounting systems; he will have standards of many kinds. He also will have a continual system of testing and working-over what he is doing to find out whether what he is doing is the best that can be done with the money he has at hand, and whether the output is as large as he could well expect with the energy and money he has invested in it. Why should we not run the schools on the same scientific basis?

Answers to such questions as what a 12-year-old boy in the seventh grade should do with a ten-figure-column addition problem have been, until recently, a matter of opinion, and usually of conflicting opinion at that. All would agree that a boy of 12 should do better than a boy of 10, but we have had no adequate conception until recently as to *how much better* he should do. We have had no objective standards of what a child should do at any specified age. With the advent of educational measurements we now are able to show what the best child will do, what the poorest child will do, and what per cent of children will be able to make a particular score.

Biology, physics, chemistry, and psychology have grown by methods of analysis and measurements. Education must imitate this prudence if she would command the respect of scientific men. To make education effective and efficient we must have an eye for causes of inefficiencies. We must discover the causes of success in some cases and failure in others. Educational measurements offer impartial and impersonal evidence which cannot be refuted.

Relation of Time Consumed to the Finished Product.

—In any well-regulated factory or business enterprise of any kind there is a definite relation between the time consumed and the amount of work done. The time it takes to make the various parts of an automobile, or a wheelbarrow, or a steam shovel, is known within quite definite limits, and an employee is quickly brought to account when this relation is seriously disturbed. How different it is in the school business. The researches made in education 25 years ago by Dr. J. M. Rice, shocked the educational world by revealing the terrible state of ignorance among educators of some of the simplest things in their profession. They tried to laugh him out of court; to treat him as an eccentric person trying to perform the impossible by measuring things in human nature. Dr. Rice's own account of the reception given him at the national meeting of superintendents at Indianapolis, in February, 1897, is both interesting and illuminating, and when contrasted with our present attitude toward him, gives us much encouragement because of the progress made. Dr. Rice had published in the *Forum* of December, 1896, an article entitled, "Obstacles to Rational Educational Reforms." In reference to this article he says:⁵

In a way that I had not anticipated I brought it directly to the notice of the Department of Superintendence at its annual

⁵ *Scientific Management in Education*, pp. 17-18.

meeting in Indianapolis, in February, 1897. I had been invited to conduct a round-table discussion on the three R's, and had expected a handful of people to talk the matter over quietly and leisurely. But it so happened that the round-table turned out to be a mass meeting, including the picked educational people of the country. After a few opening remarks I endeavored to arouse discussion on the question which I stated somewhat as follows: In some cities ten minutes a day are devoted to spelling for eight years; in others, forty. Now how can we tell at the end of eight years whether the children who have had forty minutes are better spellers than those who have had only ten?

I had expected in this way to draw out the ideas of those who believed in much teaching of spelling and those who believed in little of it, and thus to labor for a compromise; but to my great surprise, the question threw consternation into the camp. The first speaker to respond was a very popular professor of psychology engaged in training teachers in the West. He said, in effect, that the question was one that could never be answered; and gave me a severe drubbing for taking up the time of such an important body of educators in asking them silly questions.

The next speaker was a prominent superintendent, who did not like the way I had been treated and tried to come to my rescue. After this quite a number took the platform in response to calls from the audience and spoke on the spelling in a general way; but no one attempted to answer the question.

It is interesting to note that when the same association of superintendents met fifteen years later in St. Louis they devoted forty-eight addresses to the subject of educational measurements.

We naturally should expect that the ordinary principles of arithmetic would operate in finding the relation between the time spent and the results obtained. We would expect that if twice as much time is spent on a subject in one school as in another, the ratio of the finished products would bear some such relation as two to one. Such, however, was not the finding of Dr. Rice. He found classes

to whom spelling was taught incidentally just as efficient as those who had forty minutes of daily drill.

The first forward step in a problem of this kind is to get the facts on what is actually being done. Then by a comparison of the things accomplished in the various schools, tentative standards may be set up. The goals reached by the best schools may be taken as the standard of efficiency toward which all schools should work. When standards are set and facts are gathered on the standing of any school, it becomes a comparatively easy matter to measure progress.

In the absence of facts on what ought to be accomplished and what is actually being accomplished people have had to guess as to the efficiency of the school systems. If, in the opinion of school officials or the public, the schools are deficient in some particular phase, the usual procedure is to give more time to that part of the work, even though the supposed weakness may actually be the strongest link in the chain of processes. A superintendent thus confronted must do something to meet the onslaught of criticism directed against his school. Not knowing a better thing to do, he adds a little more time to that phase of the work that has been pronounced inefficient, and trusts to luck that what he has done will satisfy the public and do no harm to the school.

The following taken from the preface of the Salt Lake City Survey⁶ illustrates the point in question:

During two or three years preceding 1915 a certain amount of general criticism developed in Salt Lake City with reference to the work of the schools and the efficiency of the instruction and supervision. The harmonious coöperation which had previously existed between the Board of Education and the Superintendent of Instruction came to be somewhat impaired and the confidence of the citizens in their schools somewhat shaken.

⁶ Published by World Book Co., Yonkers-on-the-Hudson, N. Y. Used by permission of the publishers.

In particular, the rather common complaint was raised that the administration of the schools was not efficient and that the instruction in the fundamental school subjects was not producing the best results. The superintending authorities did what they could to meet such criticism by increasing time allowances and similar measures but without appreciable results. Finally the superintendent of the schools for the city recommended to the Board of Education that a survey be made.

The survey showed that the Salt Lake City schools were strongest in the very phases they had been considered weakest by the public.

An opinion circulated in the educational field soon becomes the equivalent of a law. For instance, the idea was circulated that spelling was a matter of heredity and the heirs to this wonderful faculty were very largely confined to the upper classes, yet Dr. Rice found in giving his spelling tests that the highest grades made by any children tested were made by children whose parents were Bohemian cigarmakers.

In arithmetic he found the children in the slums of some cities doing a great deal better than those from the best districts in others.⁷

The Public Is Asking for a Ledger Account of Our Business.—When the attitude of the public is impartially considered, one must come to the conclusion that it has been both tolerant and liberal. The debit side of the ledger shows the time and money expended, the number of teachers hired, the books and apparatus provided, the buildings built, and the courses of study made. On the credit side is recorded in vague, meaningless adjectives, and numbers which follow no known rules of mathematics, the standards reached and the progress made. The public, and even the teachers themselves, do not know what "excellent" means in history, what "A" means in algebra, or what "95 per cent" means in Latin. In the case of

⁷ *Ibid.*

Latin, for instance, a father would have a pretty hard time telling the relative amount of Latin his two sons knew if one were given a grade of 90 per cent and the other a grade of 45 per cent. Applying the ordinary rules of arithmetic he might expect one to know twice as much Latin as the other, but such probably is not the case. A boy may take home a grade of 90 per cent in arithmetic or a grade of 70 per cent in penmanship year after year, and when his parents sign his monthly report card they do not feel very much of a thrill over the knowledge they have gained. With practically no standards by which to work an accountant would have a pretty hard time auditing the books of a system of public schools and enlightening the public as to their efficiency.

In the school business we have a state monopoly in which there is almost no accounting. We have compelled the children to come to school. We have made courses of study which we felt were good but which were very often based on conditions past, rather than on present and future needs. After the instruction has been given to the children we have been content to trust to the growth processes to bring out the kind of development hoped for. Professor Cubberley likens the methods employed in school to the old-fashioned luck farming, where the farmer looked at the moon, guessed at the weather, put in his crop, and prayed to the Lord to pull him through another season.⁸

Inability to Show Facts Has Worked a Hardship on the Teaching Profession.—No one knows the number of good teachers who have lost their positions because they did not have facts—hard, cold, tangible facts—to prove to the public and the school officials that their work was meritorious. They have been dismissed many times because

⁸ Ellwood P. Cubberley, "The Significance of Educational Measurements," Third Annual Conference on Educational Measurements, Bulletin No. 6 (Indiana University, 1917), p. 7.

the demands were unjust. Unless a teacher has a correct idea as to where her pupils are, educationally, when they come to her, and what their native capacities are, how can she or the public or the school officials know what progress they have made? A supervisor may go into a grade room and say to the teacher, "You are doing poor work in reading." The teacher may respond by saying, "No, I am not. I am doing *good* work in the subject of reading." In the absence of a measuring stick or a standard of good reading how is one to tell who is right? It is simply the supervisor's opinion against that of the teacher. If, however, some standard has been set such as that made by Professor Courtis, the teacher may say, "Let us measure the pupils and see." If they can read with the speed and comprehension set as the standard, the supervisor will have to withdraw his statement and the teacher is sustained. On the other hand, if the class is not up to the standard, the test will quickly show it. The test being impartial and unbiased, the teacher can have no complaint to make against the supervisor for unjust demands.

As soon as school officials recognize the fact that measurements define for them just how much may reasonably be demanded, they will be unafraid of measurements. They will learn the administrative lesson that it is better to know for purposes of ordinary routine what ought to be demanded, than merely to guess at conditions.

The business man used to be offended if anyone criticized his methods or commented on his results. Now he knows that his best friend is the man who comes and tells him exactly where he stands. The one thing a successful business man cannot approve of to-day is ignorance about the results of his business. He does not fool himself any more, come what will of the revelation.

The school principal who knows in advance where his

school is weak and where it is strong is armed against criticism. More than that, he is guided in his future efforts. Some superintendents have feared the facts in reference to their schools because they knew that the number of imperfections revealed in the survey would be appalling. One has considerable hesitancy in going to the dentist knowing that his teeth are sure to "go to pieces" under the keen scrutiny of the expert. Yet his better judgment tells him to go. In the presence of a popular demand for the revelation of the imperfections and the absolute certainty that imperfections exist, it is not difficult to understand why there is a tendency on the part of many school officials to combat the movement towards widespread measurements.

The demand for measurements is likely to be especially keen if there is some parent in the community who does not like the superintendent or the principal. Such a parent may desire to "show up" the inefficiency of these officials. He never believes for one moment that the responsibility for unsatisfactory school work may be traced to the native limitations of his child or to the home atmosphere in which he grows up. Such a parent is sure that measurements will detect at some point a lack of perfection that will give his dislike for the school officer the sanction of science.

In spite of the imperfections, the American people are willing to pay and pay well for any reasonable project in education. Their liberality cannot be questioned even under present conditions. They may be made to go to almost any limit if given the facts.

IV. ESTABLISHMENT OF DEFINITE STANDARDS

It is obvious that one must have some kind of a standard before anything may be judged. If one says he has a good suit of clothes, he is basing his judgment on some

standard or standards. It may be good because it will wear well, or because it fits perfectly, or because it holds its shape well when pressed. The merchant in selling the suit may have called attention to the color; that it will not fade; that it is all wool; that the weave is of the latest design, etc. The purchaser takes the data submitted by the merchant and measures the facts by his standard of a good suit. He knows that a good suit will wear well; that it must be all wool; hold its shape well; fit perfectly, and be pleasing as to colors. There will always be differences of opinion as to which of these qualities should have the most weight. For instance, two suits may be of the same price, but one is made of better material than the other, while the second may be more pleasing as to color. One customer may prefer to take the poorer quality of goods in order to get the proper colors or weave, while another may place practically all the emphasis on the wearing qualities of the garment. Experience has taught the purchaser what his standards should be.

When we make a similar comparison to a school system, we find conditions decidedly changed. In the first place, the school official cannot furnish the individual about to judge a school system with the facts as the merchant could the purchaser of the suit of clothes. In the second place, he has not had, until recently, any way of knowing what a school system, or a particular grade in a system, ought to do. Suppose he were told the sixth grade could read a certain selection at the rate of 150 words per minute and could reproduce 80 per cent of the ideas found in it. Would he consider this a model class as to reading, a mediocre class, or a poor one? Or suppose he were told that in a certain school system the multiplication tables were taught to a beginning fourth-grade class in twenty lessons so that 95 per cent of all the children in that grade knew them perfectly. Would this be considered good,

mediocre, or poor as far as time is concerned? In other words what is the standard as to time in teaching the multiplication table?

A bricklayer has a standard and knows how long it ought to take to lay 1,000 bricks. A man shingling or lathing a house knows how long it ought to take to cover 100 square feet of surface, or how long it takes to nail on 1,000 shingles or laths. All such work is broken up into definite units of accomplishment, and standards are set for each unit.

There are hundreds of things in education that might be broken up into definite units of accomplishment, and standards of achievement set for each unit. In our country, where elementary education is characterized by the absence of system, it is not unusual for individuals, whether educators or laymen, to examine a class with a set of questions selected in an arbitrary way and judge by the results whether or not the teacher has done satisfactory work. So long as we have no definite standards, judgments based on the results of an examination may continue to do a gross injustice in estimating both the qualifications of the teachers and the value of the methods employed by them.

Three Kinds of Standards Needed.—At least three kinds of standards are needed in education: (1) standards as to quantity; (2) standards as to time; (3) standards as to quality.⁹

1. *Standards as to Quantity.*—The first step toward placing elementary education on a scientific basis must necessarily lie in determining what results reasonably might be expected at the end of a given period of instruction. Of course, the first thing that determines the standard must be the degree of capacity that the average child has for

⁹ W. W. Black, "The Movement for Greater Economy in Education," Second Annual Conference on Educational Measurements, Bulletin No. 11 (Indiana University, 1915), pp. 7-12.

the type of training that we want to give him. We cannot fix the standards irrespective of the child. The theoretical ideal of perfection must be overthrown and rational demands made, not on what an ideal child should do, but on what the average child, as he comes to the teacher, is expected to do. Then we can venture to tell the parents with assurance that their children in the fifth grade, for instance, are as good as the average even if they misspell 50 per cent of the words in a certain spelling test.

The standard set will be subject to many conditions and will have to be justified from many points of view. It will be open to question from the point of view of economy, organization, and method. Because of the interrelation of these different factors in determining efficiency, such questions as the following arise: Does the standard demand too much or too little in time and energy? Could the individual, under a different organization and method, meet the requirements of the standard more economically? Could he, under a different method, raise the standard with the same expenditure of time and energy? Is it desirable to raise the standard? Is the method such that the pupil has developed a sufficient degree of ability in applying the knowledge indicated in the standard?

Definite standards should be required in determining and checking the applications of the principles of method. When our standards as to quantity are determined, the teacher can tell exactly when the child's mechanical work is completed. This would enable him to determine the amount and character of drill work to be done.

It is not what the teacher does that counts. It is what the child does and thinks, and, until our work is organized to take advantage of these factors, we cannot hope to improve the efficiency very much. We must know exactly what is meant by "satisfactory results," then no time would be wasted when the goal is reached. Many pupils

have been kept at writing and other mechanical drills long after *satisfactory results* have been reached simply because the teacher did not know when the pupil arrived. There is a great deal, too, to be said for definite standards for the psychological effect they have on both the pupil and the teacher. Each likes to know how well he is doing and how near he is to the goal.

It may be argued that it would be impossible to secure a definite standard for measuring results generally applicable in our country on the ground that the needs of our people vary in different localities. While this sentiment deserves recognition, it will become apparent, during the course of this chapter, that proper attention to local conditions in the conduct of our elementary schools would not tend in the least to alter the measurement plan as a whole.

2. *Standards as to Time.*—When our standards as to quantity have been determined, our attention may then be directed toward the discovery of short-cuts in educational processes that will save the child's time. All would agree that we should devote a reasonable amount of time to get a reasonable result. But what constitutes a "reasonable time" is the unsolved problem. To arrive at a conclusion in this matter we must find how much time has been given to a subject in the best schools where reasonable results have been obtained and make our calculations accordingly.

Because of a lack of standards, we no doubt have expected far too much of pupils on some occasions, and on other occasions they could have done twice as much very easily. William James has pointed out that the average man lives far within his limits and possesses powers of various sorts which he habitually fails to use. There is no doubt that the requirements of school children are within their limits in many instances, and that they

could easily do two or three times the work they now do in the same time.

3. *Standards as to Quality.*—Much already has been said about quality; so a short discussion will suffice. When a child can write a specimen of handwriting equivalent to Quality 13 or 14 on the Thorndike Scale, for instance, he is said by competent judges to write well enough for all practical purposes. Any further time spent in improving quality above this point is a questionable procedure unless it is definitely known that the student will be called upon to do clerical work that may require a higher standard. The point is, that standards give focus and direction to the work and fix a point above which drill may not be profitable. Many boys and girls in school whose handwriting is equivalent to Qualities 14 to 17 on the Thorndike Scale are required to practice daily when their arithmetic and grammar need their time much more.

Standards Changing.—Some object to standards on the ground that they are constantly changing and for that reason are really not standards at all. The argument is poor, however, because that is exactly what we should expect if progress is to be made. That is what happens in business life everywhere. While standards do change, yet they are stable enough to justify their determination and are absolutely essential to the best school work. As a matter of fact, all teachers have them but they are subjective and for that reason are not as valuable as they ought to be.

V. PROFITING BY THE EXPERIENCE OF THOSE IN OTHER OCCUPATIONS AND PROFESSIONS

In the manufacturing world inventions and improvements are usually made in one of two ways: A manufacturing firm may "go to school to its competitor" and learn new ways of doing things, or it may offer a reward

to its employees who can devise new and better ways of doing things. Any patent which is a labor-saving device is quickly appropriated by men in all lines of business. There is no hesitancy in appropriating new ideas irrespective of their source, the only question being, Will they work? In the recognized field of science, such as physics, chemistry, medicine, etc., the members of the profession are not only willing to learn from each other but they are compelled to do so under penalty of law. Those who fail in practice to give due recognition to important discoveries are held responsible for the consequences.

If education is to become a science it must utilize scientific methods just as other businesses and professions have done. We must discover some truths in regard to educational processes which, if ignored by the teacher, will make him liable to prosecution for malpractice just as the physician who has bungled the setting of a bone.

Teachers make mistakes year after year simply because no record is made of their procedure and because there are no signposts to guide the new ones away from the pitfalls. Successful businesses and professions are not operated in such a haphazard manner.

Our government maintains a Bureau of Standards at Washington with exact units for all measures of mass, length, and time, with a large number of derived units. For all current work in physical science these standard units are essential and without them modern physics and chemistry could not exist. Every new discovery or invention in science must be stated in terms of them, and every physical and commercial enterprise in modern civilization is absolutely dependent on them. It is not surprising that a physicist with highly refined instruments for physical measurements should look doubtfully upon the yet immature efforts of measuring mental qualities and mental achievements. Would it be possible in education

to get the same kind of measurements we have in physical science, as measurements in length and weight? We can tell a boy's stature in a perfectly definite way so that everybody in the world will know what we mean. We can tell the weight of a person or the number of pounds he can lift, and the world accepts our measures understandingly and without question because the units have been standardized.

The following examples from the business world indicate the great painstaking care that successful business takes to see that efficiency is maintained throughout its system, and how responsibility is placed on each one in the system to carry out his particular part of the programme.

In the Middle West a great corporation manufactures stoves in large quantities. The shop history of each stove is kept in detail from raw material to the shipment of the finished product. The name of each workman responsibly concerned in the inspection of the raw material, in the making of any part, in the examination and shipment of each part, is recorded. A complaint from the purchaser immediately locates the responsibility upon the workman at fault. The article upon which all this painstaking care is lavished sells from fifteen dollars up. The method of successful business makes it possible to place responsibility where it belongs. The public schools might well imitate this prudence.

The Manufacturer Has Specifications.—The manufacturer knows exactly what he is trying to do. He draws his specifications according to the needs of the market he is trying to supply. He knows how much of his output is efficiently manufactured and how much raw material goes to waste in the dump pile. His object is to obtain a more satisfactory output so there will be a larger profit for the business.

The ideals and processes of scientific method in educa-

tion are in salient respects similar to those that are reshaping the processes of industry. In education as in industry the scientific idea is, at base, analytic scrutiny, exact measuring, careful recording, and judgment on the basis of observed facts.¹⁰

The school principal who tests his school is guided in his future efforts. The fact that adverse criticism causes no shock is of some importance, and the positive result is that the school is stimulated to improve itself. We have awakened to a startled realization that, in education as in other forms of organized activity, applied science will do the work much better than the old trial and error methods. Even those processes that have rested secure in the sanction of generations may be improved by the application of science.

In dealing with the application of scientific methods to education and to industry we must ever bear in mind two fundamental distinctions. These relate to the use of time and the types of product in these two kinds of activities. In industry the finished product conforms to a definite pattern. It is a constant. The variable is the time. If the finished product is to be a wheelbarrow of a certain design, the problem is, How long will it take to turn out a wheelbarrow according to the specifications laid down in the pattern? The task to be done is always definite. In education it is different. The time is a constant, eight years in the grades, for instance. There is no definite pattern according to which we try to mould the lives of the boys and girls. Their native capacities, aptitudes, and proclivities are different. We expect training to develop more differences than fewer. These fundamental differences must be kept in mind when business methods and education are compared.¹¹

¹⁰ Leonard P. Ayres, "Making Education Definite," *ibid.*, p. 87.

¹¹ *Ibid.*, pp. 87-88.

VI. MEETING SOME OBJECTIONS TO EDUCATIONAL TESTS
AND MEASUREMENTS

In launching any great campaign for reform in method and ways of doing things, one must expect considerable opposition from not only those engaged in the business but from outsiders as well. Meeting the arguments against such movements becomes, therefore, as vital a factor in getting the new methods in operation as some of those which are very much closer to the real problems to be solved. In this part of the discussion we shall note some of the objections which have been raised in reference to tests and measurements in education.

Educational measurement, like all new movements, has its scoffers and its zealots. It is, of course, easy to point out imminent dangers in measurements. Like all new movements it will have to run the gauntlet of criticism. Measurement in education presupposes commensurable quantities, yet it is known that many educational products are incommensurable. There are spiritual and material products in education. Over the former there will ever be the veil of mystery and doubt. Man spiritually has always been the enigma of existence and will continue to be so. We can never plot the curve of genius nor measure the unit of inspiration. But there are a vast number of educational products that are measurable. The day of the educational engineer is at hand. The day of educational opinion will go, and the sway of dominant personalities in education will be limited by facts.

When a new movement is launched, the best thing to do is not to endorse it, or condemn it, but to study and understand it. This is not what always happened in education, however. But the opposition to change is not peculiar to education. It makes its appearance in other professions and industries. The important thing is to be

able to meet the objectors with hard, cold, irrefutable facts that will prove their contentions to be wrong. The objections to tests are many and varied. We shall note a few of the more common ones.

1. *Tests will not endure.*—Some opposing the testing movement say that tests will not endure the ravages of time and change; that they are of temporary value, and those we are using to-day will be "scrapped" for something else to-morrow. In answer to this objection all we can say is that we do not expect them to endure in their present imperfect form. Progress is made by casting aside the outgrown and imperfect tools for the more modern instruments. Other sciences have progressed in this way. The pseudo-sciences of ancient times supplied the foundations for the later scientific achievements. In any advancing society, old knowledge, old philosophies, and old culture must constantly be in a state of reconstruction to keep pace with the race's progress.

2. *The child mind is too complicated to measure.*—There are those who say that the mind is complicated by so many elements that enter into its development that no definite conclusions can be drawn. They are supported in this view by the fact that even broad-minded teachers of wide experience differ on the most elementary points coming under their daily observation. And this further item may be mentioned in their favor, that even the same teachers are constantly changing their views; they no longer believe in one year what they firmly believed the year before; and a year later they will begin to feel that their second theory was wrong and their first was right, and so on indefinitely. This state of affairs comes about because judgments have not been based on facts but on mere opinions. Whatever the reasons may be for changing their opinions, educational measurements tend to stabilize them because they are based on facts.

Some teachers, for instance, will tell us we cannot measure thought progression in English composition, nevertheless those same teachers mark one composition 75 per cent, another, 77 per cent, and a third, 74 per cent. When they have had 75 per cent as a passing mark they have continually refused to promote those who were given a grade of 74.

3. *The judgment of a competent man is better than scales.*—A strong objection is made to scales on the ground that the common-sense judgment of a first-rate man is better without these units and scales than the action of a stupid or incompetent man with them. The important thing is not whether a first-rate man can do better without such scales than an incompetent one with them, but rather that the efficiency of each is increased. On the other hand, it is precisely the work of science to get good work done by those who are rather mediocre. Thanks to the progress of science, we can now solve problems that Aristotle could not. We would all prefer to have a stupid doctor of to-day who, nevertheless, understood the use of antiseptics and antitoxins than Galen or Hippocrates, though in respect to common sense there would be no choice.

The dangers of measurements are as real and imminent as the advantages are self-evident. Dangers will arise from the mass of superficial and erroneous results that will certainly be presented to the educational world in the guise of scientific contributions applied to pedagogy. But we must welcome all these contributions, challenge them, and attempt to verify them. The sciences cannot escape injury if their results are forced into the rush of the day before the fundamental ideas have been cleared up and an ample supply of facts collected.

4. *Tests tend to reduce all educational work to a dead level with no allowance for individual differences.*—As to

the tendency to obtain uniform results, the effect is meant to be just the opposite to uniform. Our surveys are giving us a knowledge of conditions, both in school and out of school, under which the individual child lives. Physical tests will tell us more accurately than we have been able to know heretofore the ability of each individual. With this accurate knowledge of the individual child, there is no way open but to make his self-furnished standards the guides in his case. This charge of reducing all individuals to a dead level is born of a complete misunderstanding of the aims and processes of the new method. Its aim is not uniformity but individual development. People fear lest educational experiments will make children lose their spontaneity, or spirituality, or something of the sort. They feel a certain skepticism about tests and scales. We should not care much for the teacher who did not have a great interest in the finer, subtler qualities of character and tastes. At the same time an artist can preserve all his interest in the finer, subtler qualities of taste and still use the compass points to measure his distances as he draws. "Certainly, that old rebuke made in the first days of child-study movement that if we were at all scientific with our children we would come to love them less is out of place. It is the mothers who love their babies most who weigh them oftenest."¹² The mother who weighs her baby every month, and to an ounce, is in these days precisely the mother who loves her baby on the average as well, if not better, than others.

Teachers must not expect tests to do everything for their schools. The rather mediocre intellect of youth which repels information will not blossom out overnight by some

¹² Edward L. Thorndike, "Units and Scales for Measuring Educational Products," *Proceedings of a Conference on Educational Measurements*, Bulletin No. 10 (Indiana University), pp. 128-141.

sort of hocus-pocus of measurements into a marvelous genius who will learn anything set before him. A part of the trouble has been the fact that our specifications have not always been dictated by the needs of the future. Too often they have been framed by those who are thinking in terms of the past. We wish to train the children in our schools that they may earn efficiently and comfortably their daily bread, but most of us want them to become more than skilled bricklayers, carpenters, ironworkers, stenographers, telegraphers, more than capable physicians, lawyers, and merchants. If practical efficiency is all the American schools can do for a child, they are certainly far from efficient.

Education will, of course, always need its poets, its artists, and craftsmen, as well as its managers and men of science, but it needs these also. There is no reason why the artistic life should be impeded by measurements. Of course, we cannot subject all human life to mathematically accurate tests, certainly not in the near future. Human life is deeper than all our mathematics.

This agitation for better-measured products is not confined to education. Commerce, industry, politics, philanthropy, and other great fields of endeavor, are similarly agitated. Before the impact of oncoming generations, long-worshiped and long-cherished idols are falling. The old order yields slowly. It rejects the new as propaganda. Thus the tide of conflict moves back and forth in the great fields of human endeavor, disclosing in the lull of struggle the inevitable process of change.

5. *Tests measure so small a part of intellectual life that they are not indicative of general ability.*—An educational product is usually complex, and its measurement is as difficult as measuring an elephant. There are many characteristics to be taken into consideration. We do not make a complete measurement of the total fact, but we measure

the amount of some feature. Every measurement represents a highly abstract and partial treatment of the product. Many critics who object to tests and scales do not understand this. An educational product invites hundreds of measurements. In making an automobile each part is measured in its own peculiar way. Linear measure is used for length, cubic measure, for volume, and the strength of the steel is tested by another unit. In exactly the same way the traits, characteristics, mannerisms, powers, and skills of an individual would have to be subjected to many kinds of measurements before their strength and quantity could be determined. If a teacher is a good one, she is good because of the presence or absence of certain powers, traits, skills, and mannerisms, or she is good because she has these characteristics in a certain proportion. Dr. Thorndike has pointed out the fact that anything that is, exists in some quantity. If it exists in some quantity, it can be measured. Perhaps we cannot measure it to-day but we may be able to measure it some day.

People have said, and rightly, that we can never determine mathematically the degree to which a strong man and a noble woman influence for good the character of the pupils. But they overlook the fundamental truth that in education, as in other pursuits of life, character and efficiency go hand in hand. As school executives make practical application of the newer scientific tests, no fact stands out with more impressive distinction than that the teachers whose classes make the best record are the teachers who are the most truly successful in the shaping of character.

VII. KEEPING AN ACCURATE RECORD OF ALL METHODS TRIED AND PROGRESS MADE

Measuring progress implies a starting point and a goal. The significance of this fact should not be overlooked in

education. Attempting to measure progress in school work, when either the starting point or the goal is unknown, would be like attempting to measure the progress of one "joyriding" about a city with no definite goal in mind. All that could be said is that he had been out one hour, two hours, or some other length of time. If, however, one were driving an automobile from Portland to Seattle, it becomes an easy matter to measure progress, because there is a definite starting point and goal. A definite point of departure and a definite goal is also needed in education if progress is to be measured.

VIII. THE CULTIVATION OF THE CONFIDENCE AND THE UTILIZATION OF THE SUPPORT OF THE PUBLIC

A new kind of confidence must be created on the part of the public in our schools. This confidence constitutes the capital with which the efficient school system must develop its dividends and activities. In devising educational procedure, therefore, we must constantly have in mind the intelligent and public-minded portion of our citizenship. As education grows scientific it tends to become less intelligible to the public. With a growing technical terminology the educational thinker tends to speak a dialect difficult for the ordinary person to comprehend. The result is that as the education has become more scientific it has tended to isolate itself from the understanding of the people.

Teachers Must Know Why Sweeping Changes Are Made.—It has been said that there is a growing tendency for some city superintendents in large and somewhat centralized systems of public schools to make sweeping changes in schoolroom procedure without consulting, and what is more serious, without attempting to convince their teachers of the necessity for such changes. There can be no true profession of teaching where most of the members are

required to carry out official orders mechanically and blindly. A clear understanding of underlying principles is essential to good teaching. In the art of administering to the intellectual and moral awakening of childhood, the spiritual worker should fully comprehend the meaning of the plan and should have a familiarity with the tools with which he works. Successful educational leaders are realizing more and more the necessity of carrying their teaching staff with them in all progressive reforms, not simply as a matter of respect for the teachers, but as a matter of necessity in getting the essential work of the school done.

It is necessary that the thought of the leaders be transmitted to the rank and file, to all trainers of youth, to parents as well as teachers. Our educational institutions are not made by imperial edicts and bureaucratic decrees handed down from educational experts. They must come from the people. It is not sufficient that educational leaders alone should know the significance of a given reform or movement. The public must understand and accept the proposed policy. The intellectual channels between leaders and followers, profession and populace, must be kept open. Both teachers and public must know the limitations of the school. The leaders can advance no farther with an educational movement than the public will endorse. Tests and measurements, however technical in character, must not have their significance clouded. The public wants to understand its educational system and what it is supposed to do. The ultimate worth of the principles and devices must be measured by the extent to which they are within the comprehension of the typical layman. Our aim must be the bulwarking of the cause of public education by a common confidence in the schools. This new confidence cannot be sustained on vague theories nor upon standards that are established in defiance to either scientific procedure or common sense.

The Public Is Interested in Education.—If our schools are to have credit for their work, this new confidence must obtain. It is true that people as a whole are not interested in pedagogy because they do not understand it, and they are not in sympathy with the pedagogues because they do not understand their subtle minds. But they are intensely interested in education. It is a mistake to think that the lack of educational progress may be attributed to public indifference and its consequences. The people are willing to dip down into their pockets to almost any depth with reverence and, as a rule, without the slightest murmur.

The teaching profession has done but little to demonstrate to the public that what it was doing was worth while. Results have always been recorded in vague, intangible adjectives such that the public could not understand. That the public has never taken an intelligent interest in its schools is not its fault, but that of the educators themselves. For how can the public be expected to distinguish the true from the false when the leaders in the profession do not agree on the simplest fundamentals?

The ultimate cause of the lamentably slow progress toward the introduction of educational reforms may be traced, therefore, beyond the province of the general public into the professional circle itself to an inner strife and turmoil consequent upon the uncertainties in which the entire problem of elementary education is involved.

The so-called practical man is especially insistent in his demands for tangible results.

With our methods of reporting results it is really wonderful how liberally the public has contributed to the support of the public schools and how little the adverse criticism has been. So liberal has the public been that to-day we are spending more on education than all the rest of the world. In criticism of our work the business

man does not say that the child does not know anything at all about adding, for instance, but he says he does not add with the proper speed and accuracy. He does not say the child cannot understand English at all, he says he is not intelligent and cannot follow instructions. He does not say the child cannot read, he says he reads too slowly and misunderstands what he reads. Why should we not, when we are dealing with purely mechanical skills like the above, give absolutely definite specifications of what we expect and see to it that no child is marked "passed" until these definite specifications are reached?

The American people has had an abiding faith that the great public schools would somehow eventually push everyone to the place for which his disposition, talents, and psychophysical gifts, prepare him; that they would discover the strong and weak points of the children who will some day occupy the places of responsibility. With that thought in mind, and without demanding a strict accounting for the vast sums so lavishly spent, they continue to have faith in the schools the accounts of which they are unable to audit because only the debit side of the ledger is kept. Science has changed other professions and businesses so that the public is becoming accustomed to look for tangible results. Just as in the factory when a piece of raw material is put through a certain process, we look for tangible, measurable results, so in the school system, when we put a child through an educational process, we look immediately for measurable results. If the desired result is still lacking, then the process may be repeated or a new process or method tried. This is undoubtedly the only sane way to teach school. The old method was to put the child through the process and, without measuring the result, trust to luck that the desired goal was reached. If the boy became a man of prominence,

the school-teacher was liable to take considerable credit to herself for teaching him correctly.

In this modern intensive life we cannot run the risk in waiting until adult life is reached to find out whether our educational processes are bringing the proper results. It is then too late. Society demands that we find out immediately what changes are wrought by our educational processes. This information may be gained only by having definite units of accomplishment and scales and units of measurement. The public demands that the results of training be recorded in sensible scientific units and that they be given immediately after the child has been put through the educational process so that deficiencies may be provided for before the child has passed beyond the influence of the schools.

Fields of Educational Tests and Measurements.—There are many fields into which the subject of educational tests and measurements may be divided. Seven are perhaps definite enough to deserve special mention. These are: (1) the measurements of intelligence; (2) the measurements of school achievement; (3) the measurements of the materials of instruction; (4) the measurements of the physical growth of school children; (5) the measurements of the money cost of education; (6) the measurements of school buildings; and (7) the measurements of retardation, acceleration, and elimination.

Each of these fields is subdivided into narrower ones, and some are re-combined to produce other fields as, for instance, when measures of intelligence and school achievements are pressed into service to determine, in part, the efficiency of a teacher. Space will not permit an exhaustive treatment of these fields in an elementary work of this kind, any one of which furnishes sufficient subject matter for lengthy analyses and discussions. Nevertheless, a somewhat lengthy discussion of the first two fields mentioned

above will be given because it is in these fields particularly that the regular classroom teacher should acquaint herself with the tools for measuring. It is in these fields that measurements come in direct contact with the learning process in the regular classroom work. The other fields, especially the last three, perhaps pertain more to the administrative phases of education, and, while they affect the regular room-teacher, they are not quite so intimately connected with the regular recitation work. Five chapters will be devoted to a discussion of the first two fields, while the other five fields will be discussed in a single chapter. The dominant idea in this single chapter will be to map out and orientate the various divisions rather than to treat them exhaustively. In fact, the discussion will be limited to a few of the most characteristic phases of each of these five fields.

BIBLIOGRAPHY

1. AYRES, LEONARD P., "Making Education Definite," Second Annual Conference on Educational Measurements, Bulletin No. 11 (Indiana University, 1915), pp. 85-96.
2. AYRES, LEONARD P., "The Measurement of Educational Processes and Products," *ibid.*, pp. 127-135.
3. BLACK, W. W., "The Movement for Greater Economy in Education," *ibid.*, pp. 7-12.
4. BOBBITT, JOHN FRANKLIN, *The Curriculum* (Houghton Mifflin Co., 1918).
5. BUCKINGHAM, B. R., "Efficiency Indices," Third Annual Conference on Educational Measurements, Bulletin No. 6 (Indiana University, 1916), pp. 85-118.
6. BRYAN, WILLIAM LOWE, "Common Sense and Science in Education," Proceedings of a Conference on Educational Measurements, Bulletin No. 10 (Indiana University, 1914), pp. 8-9.
7. BAGLEY, W. C., "The Determination of Minimum Essentials in Elementary Geography and History," National Society for the Study of Education, *Fourteenth Year Book*, Part I, pp. 131-146.

8. CUBBERLEY, ELLWOOD P., "The Significance of Educational Measurements," Third Annual Conference on Educational Measurements, Bulletin No. 6 (Indiana University, 1916), pp. 6-20.

9. CHARTERS, W. W., "Scientific Curriculum Construction," Sixth Annual Conference on Educational Measurements, Bulletin No. 1 (Indiana University, 1919), pp. 78-94.

10. MONROE, WALTER S., "The Next Step in Educational Measurements," *ibid.*, pp. 94-103.

11. RICE, J. M., *Scientific Management in Education* (Publishers Printing Co., New York, 1913).

12. RUSK, ROBERT R., *Introduction to Experimental Education* (Longmans, Green & Co., 1913).

13. THORNDIKE, EDWARD L., "Units and Scales for Measuring Educational Products," Proceedings of a Conference on Educational Measurements, Bulletin No. 10 (Indiana University, 1914), pp. 128-141.

CHAPTER III

THE MEASUREMENT OF INTELLIGENCE

General Statement of the Problem.—It is a platitude to say that individuals differ from one another in intellectual achievements. We have known this from time immemorial. The reasons *why* they differ from one another are known only in part. It is known that one's intellectual ability depends upon two factors: (1) native capacity, which is a physical inheritance and therefore beyond the control of the school; and (2) environmental conditions, a part of which may be directly controlled by the school. It is the purpose of this chapter to discuss some of the problems and technique that have to do with measurements of the first factor. The particular phase of this factor that we desire to measure is that known as "general intelligence." The subject is naturally divided into a number of major problems each of which is subdivided into a number of lesser ones. Some of the major problems are: (1) What is meant by general intelligence? (2) What shall be the nature of the tests to measure general intelligence? (3) What shall be the units of measurement? (4) What shall be the methods of scoring? Many lesser problems will arise as the discussion proceeds.

Before discussing these problems, a brief statement as to the work which was being done in experimental psychology and psychiatry (that branch of neurology which treats of mental disorders and of the organic changes associated with them) that led up to the attempts to

measure intelligence will help to clarify the situation and throw some light on the questions as to why this movement came into existence and why certain tests were used in the early stages of the work. Space will not permit more than the very briefest statements as to general status of psychology prior to the attempts to measure intelligence.

Until very recent times it was generally conceded that psychical phenomena could not be subjected to quantitative measurements. Real experimental psychology came into being within the memory of men now living.

The honor of founding quantitative psychology belongs to Gustav Theodor Fechner (1801-1887). It was as late as 1860 that he published his *Elemente der Psychophysic* which brought together scattered observations from astronomy, physics, and biology, which, together with his own elaborate observations in physics, mathematics, and physiology, were placed at the service of mental measurements. A student who would understand the principles and general methods of mental measurements must still go to school with Fechner.

Wilhelm Wundt (1832-1920) did more to encourage and inspire students to work in this field than any other modern scientist. He established the first psychological laboratory at Leipsic in 1878. When he began his experimental work in psychology, there was little save the psychophysical laws announced by Fechner, reaction times, the experimental physiology of the senses, and the early studies in brain localization.

At the time Wundt established his psychological laboratory, the general field of psychology was beginning to be divided into various fields each with its own methods of investigation. There were *subjective psychology* which relied wholly on inner perception, and *objective psychology* which attempted to perfect and to supplement inner perception by objective means. *Objective psychology* was

again divided into:¹ (a) *experimental or physiological psychology*, which brought inner perception under the control of experimental appliances; and (b) *social psychology*, which sought to derive general laws of psychological development from the objective products of the collective mind.

Wundt realized that if psychology was to advance it must follow the inductive path. Two inductive methods were available: (1) *The method of statistics*, which is indirect and bears primarily on the practical with little emphasis on theoretical psychology. This is the method that furnishes psychology with its facts. (2) *The method of experiment*, which, as Wundt says, is a principle applicable over the whole range of psychology. It was the latter method in which Wundt was primarily interested and the one that chiefly concerns us here.

The physiologist E. H. Weber, desiring to determine the power of the skin and of the muscle-sense to discriminate between weights, began as early as 1830 to make experiments in the field of sense perceptions. Weber devised schemes for measuring the relation between the intensity of stimuli and their corresponding sensations, which relations were designated by Fechner as *Weber's laws*.

When certain psychological tests were devised to investigate special mental functions, it was found that the results obtained from these tests varied with the degree of intelligence of the subjects. When this discovery was made, these tests were diverted from their original purpose and were employed in diagnosing endowment or general intelligence. For instance, on investigation it was found that in sensory discrimination the most intelligent

¹ E. B. Titchener, "Wilhelm Wundt," *American Journal of Psychology*, Vol. 32, Apr., 1921, pp. 161-178.

subjects had the lowest thresholds. It was concluded therefrom that all that was necessary was to measure the subject's power of sensory discrimination, such as the sensitivity to differences in pitch, or the ability to distinguish between the length of two lines, in order to determine the degree of intelligence of an individual. It was thought also that tests in memory and association would give results indicative of the degree of intelligence. Further investigation, however, proved that these assumptions were not true. It happens very frequently that an exceptionally good memory is possessed by one with low intelligence. ★

From this humble beginning there has been organized a vast body of methods and results requiring several hundred standard experiments and a great army of experts who spend most of their time in the laboratory employing these new methods of observation and introspection.

Why Work in Psychological Measurements Was Retarded.—The intrinsic difficulty of psychological measurements is undoubtedly largely responsible for their tardy advent and lack of progress. The recesses of the mind were apparently inaccessible. But there are other reasons why progress was not made in this science.

Descartes had drawn a sharp line of demarcation in popular thought between the natural and the mental sciences. The former were considered quantitative, and hence measureable, the latter, qualitative, and not measureable. This popular "common-sense" point of view has had the weight and inertia of a settled tradition. Immanuel Kant (1724-1804) reinforced the "common-sense" view from the fields of philosophy. He declared, in 1786, that psychology never could obtain the rank of pure science. To overcome the effects of a dogmatic statement of a philosopher like Kant was a difficult thing to do. Of course, there were plenty of ideas and suggestions as to what might be done. The path of scientific progress

is littered with brilliant suggestions. It is so easy to suggest but so difficult to grasp the suggestion and carry it to its conclusion.

Not only did mental measurements get a late start but their very nature makes progress slow. Like political constitutions, new branches of knowledge are not made but must grow. Improvements will be slow and discussion is one of the means by which they are made.

Effects of Wundtian Laboratories.—Just as the Binet tests have occupied the center of the stage in the field of intelligence measurements, so Wundt and Wundtian laboratories have been the center of the movements in experimental psychology for the last fifty years. Many of Wundt's students, Titchener, Hall, Kraepelin, Müller, Cattell, Meumann, and others, have gone far beyond him in many lines. Dr. Stanley Hall founded the first psychological laboratory in the United States at Johns Hopkins University in the eighties. Within a few years Wundtian laboratories in experimental psychology were established at Philadelphia and at Columbia by Cattell; at Cornell by Titchener; and at Harvard by Münsterburg.

Beginning in the small and more accessible fields of sensation, experimental methods gradually moved up toward the more complex problems of association, attention, voluntary movements, will, the higher thought processes, and even feelings. But little has been done, however, in the last-named field. Of all mental processes, the feelings, the affective states, are the most baffling and least amenable to measurements.

As early as 1912 fifteen of the more progressive asylums were applying Wundtian methods to some extent in the study of insanity. Wundt realized, as few scientists did, that progress in a science is bound up with progress in the methods of investigation. Every new instrument used is followed by a series of new discoveries, and modern

science itself originated in a revolution of methods in the hands of Bacon, Galileo, and others.

Psychology made little progress from the time of Aristotle to Wundt because there is no scientific field so crowded with presuppositions and prejudices and so barren of scientific methods of investigation. It is the spirit of Wundt more than anything else, perhaps, that has led to the recent improvement in controlled observational methods in the study of animal instinct, from tropisms up to tests of intelligence.

The methods and tools developed and employed in the psychological laboratory were soon to fall into the hands of those interested in mental development not only from the standpoint of applied science, but also with the idea of helping some of the more unfortunate ones in our society. Men motivated by altruistic and philanthropic ideas sought to better mankind by the application of science in the field of intelligence. It was the psychiatrists dealing with abnormal adults who first wanted to test intelligence. The movement was at first wholly within the field of psychopathology. These men devised tests and systems of tests. By far the greater portion of these tests took on the character of questions and qualitative tests rather than that of quantitatively gradable tests. Their methods were open to many criticisms. The method of determining intelligence was to test a few individuals a great many times with one or more tests and then apply the tests to a great many individuals only once, or rarely more than two or three times. If it was found, on the whole, that persons known to possess more than average intelligence obtained better averages than the less intelligent, it was assumed that the methods employed would answer for the testing of intelligence. Most of the early testing was of this sort. Their interest was primarily with abnormal individuals. Whatever testing the psychiatrist

did on normal individuals was simply to establish norms for the measurement of abnormal adults. Whatever data were gathered in reference to normal individuals came about as a "by-product" of this process. They knew little about standards for normal adults with which the performances of abnormal subjects were to be compared, and they knew absolutely nothing about standards for children. What is more, one normal standard is not enough. With children every age-level must have its own standard. The magnitude of a defect in intelligence in a nine-year-old child can be determined only by comparing it with a normal nine-year-old intelligence.

In the purely psychological fields men were studying the psychology of testimony, optical illusions, association experiments, tachistoscopic experiments, the learning of nonsense syllables, etc. The first tests made by psychologists were not designed as measures of intelligence. They seem to have arisen as a direct result of the individual differences noted in the laboratory by the experimental psychologists. At first these individual differences were a distinct hindrance because they made the establishment of psychological laws difficult. But psychologists became interested in them for their own sake, and once this occurred we had the birth of the test designed to measure the mental differences between individuals. The first tests were concerned with measurements of specific "faculties" or capacities. They were tests of different mental processes or of different states of consciousness. At first the tendency was to ignore individual difference of pupils. Now the demand is to individualize instruction which will accentuate these differences. The opinion is not unanimous that differences should be accentuated, but the tendency towards individual instruction will undoubtedly bring about this result.

Many persons were using these individual experiments

as tests of intelligence. The chief hindrance both to the psychiatrist and to the psychologist in attempting to measure intelligence up to this time was, perhaps, a lack of definite working hypotheses as to what intelligence really is. The ordinary school examinations would give them a notion of the pupil's knowledge and of his external accomplishments, but they do not afford an index to his inner endowment, his mental maturity and power. What this endowment is, and a means for measuring it, were obviously the next steps in the scientific procedure of mental measurements. The more or less blind probing with poor methods for an unknown characteristic brought poor results. They realized that tests must be selected in accordance with rather definite hypotheses of the exact nature of intelligence. Therefore definite hypotheses of just what intelligence is, with better and more scientific measuring instruments, were the crying needs of the hour. It is here that work began in earnest on the measurement of intelligence.

What Is General Intelligence?—The following definitions of intelligence can, of course, be nothing more than working hypotheses in this field. We shall note a number of definitions of this kind. Stern² defines intelligence as "*a general capacity of an individual consciously to adjust his thinking to new requirements: it is general mental adaptability to new problems and conditions of life.*" He thinks this definition clearly differentiates intelligence from other mental capacities such as *memory*, *genius*, *talent*, etc. Adjusting one's thinking to new requirements obliterates the effect of *memory*; forcing the examinee to adapt himself to the performance of problems already set differentiates intelligence from *genius*, the nature of which

² William Stern, *The Psychological Methods of Testing Intelligence*, translated by Guy Montrose Whipple (Warwick & York, 1914), p. 3.

is to create the new spontaneously. The fact that it is a general capacity distinguishes intelligence from *talent*, the chief characteristic of which is the limitation of efficiency in one kind of content.³

Binet's conception of intelligence emphasizes three characteristics of the thought process: (1) its tendency to take and maintain a definite direction; (2) the capacity to make adaptations for the purpose of attaining a desired end; and (3) the power of auto-criticism.⁴ In his earlier work Binet believed that *the essence of intelligence was capacity to adjust the attention*.⁵

Meumann's conception of intelligence is not quite clear. At times he lays great stress on the *understanding of the abstract* as the root of intelligence. He makes use of the retentive powers of memory in the learning of abstract words; he holds that the power of independent and creative elaboration of new products out of the material given by memory and the senses is a manifestation of intelligence. In practical affairs intelligence, according to Meumann, means the ability to avoid errors, to adjust one's self to his environment, and to surmount difficulties. He makes extensive use of what is known as the "Masselon experiment," which gives the subject a number of words that are to be used in a sentence. He thinks this is a reliable index of the maturity of the associative processes.

According to Ebbinghaus, the essence of intelligence lies in comprehending together, in a unitary meaningful whole, impressions and associations that are more or less independent, heterogeneous, or even partly contradictory. "*Intellectual ability consists in the elaboration of a whole into its worth and meaning by means of many-sided com-*

³ *Ibid.*, pp. 3-4.

⁴ "L'intelligence des imbéciles," *L'Année Psychologique*, 1909, pp. 1-147.

⁵ Stern, *op. cit.*, p. 17.

bination, correction and completion of numerous kindred associations." ⁶ He thinks that every true instance of intellectual ability may, in the last analysis, be reduced to an act of combining. It is a combination activity. To test the maturity of this combining activity, he gives the subject sentences broken up into parts with gaps in them, words left out, and he asks the subject to supply the missing parts so as to make the sentences read correctly. The same principle was used by Healy in the Picture Completion Tests. It is also used in many other experiments.

Zeihen attempts to measure intelligence by the use of tests of retention, development, comprehension, and generalization.

Is Intelligence a General Faculty of the Mind?—At the present time, the term *general intelligence* is commonly understood to mean an innate ability or group of abilities that lie at the basis of the acquired intelligence of an individual. We know that intelligence itself is not in-born, but only the capacity to become intelligent. Whether general intelligence signifies a single inborn capacity which functions in all situations, or a large number of specific capacities, more or less related, which enable an individual to acquire intelligent behavior in many different activities, is a question that has not been settled by psychologists. Spearman, Hart, and Burt explain innate intelligence as a "general common factor." Spearman has developed a mathematical formula which shows the correlation between the various faculties of the mind, and hints at a general common factor in all mental performances, which is known popularly as "general intelligence."

Burt, employing in his investigations the methods of Spearman, concludes that there is a general function—

⁶ Cf. Lewis M. Terman, *The Measurement of Intelligence* (Houghton Mifflin Co., 1916), p. 46.

a greatest common measure—permeating, to a greater or less extent, the various special functions measured by his tests. He thinks that the measurements obtained are measurements, more or less indirect, of a single capacity, and not determined purely by different capacities in different cases; that the idea of an all-around mental efficiency applicable in many directions is a legitimate conception.⁷

Thorndike does not endorse the idea of the existence of the common factor denoted by the term “general intelligence.” Relative to this point he says:⁸

This doctrine requires not only that all branches of intellectual activity be positively correlated, which is substantially true, but also that they be bound to each other in all cases by one common factor, which is false. The latter would require that no two intellectual abilities or branches of intellectual activity should be more closely related to each other than to the fundamental function by which alone they are supposed to be related. . . . But unless one arbitrarily limits the meaning of “all branches of intellectual activity” so as to exclude a majority of those so far tested, one finds traits closely related to each other but with their common element only loosely related to the common element of some other pair. . . . The mind must be regarded not as a functional unit, nor even as a collection of a few general faculties which work irrespective of particular material, but rather as a multitude of functions each of which involves content as well as form and so is related closely to only a few of its fellows, to the others with greater and greater degrees of remoteness.

Meumann and others criticize the idea of a general faculty known as general intelligence and cite many reasons why this hypothesis is contrary to the facts. The arguments are too long to present here, and suffice it to say that most psychologists deny the existence of this general faculty.

⁷ *Child-Study*, Vol. 4, pp. 97-101.

⁸ *Educational Psychology*, Vol. III, pp. 364-366.

We might go on at length giving definitions of intelligence, which are, at best, nothing more than working hypotheses; nevertheless, such hypotheses serve as guides in our attempt to measure an inheritance which conditions all our mental achievements.

Inability to Define Intelligence Accurately Does Not Prohibit Measurements.—It may seem at first thought that it would be impossible to measure a thing that has not been defined. This, however, is not the case. Stern points out that electromotive force was measured long before electric currents were well understood; that many diseases were diagnosed and successfully treated when very little was known of their real causes. The whole science of chemistry, for instance, is built up on the supposition that matter is composed of molecules and atoms, and the definitions given for them are at best but working hypotheses. So the handicap of being unable clearly to define intelligence may not be as great a hindrance as it might seem. The above definitions of intelligence differ more because of different points of view than because men do not agree as to what intelligence is. The tests that we shall now describe attempt to approach this capacity from many angles; hence the difference in the tests.

Since the Binet tests are by far the most important, both as to origin and as to the fundamental principles used, a short historical sketch of their development will throw much light on attempts at measuring intelligence. After this brief historical statement, we shall again refer to the nature of intelligence in the discussion of the types of tests used to measure it. Then we shall give some of the most modern conceptions of it as discussed by writers in this field.

The Binet Tests.—The Minister of Education in France in 1904 decided to separate the subnormal from the normal children in the public schools of that nation. For this

difficult task he called upon Alfred Binet to devise a series of tests which might be used for that purpose. Binet had had a wide experience with the education of children. He had been president of the *Société Libre pour l'Étude de l'Enfant* for a number of years. At the suggestion of many teachers he had organized a committee for the care of abnormal children which initiated various investigations relative to backward children. He had worked with children for many years studying their peculiarities and proclivities.

Binet consented to undertake the difficult task and, calling to his assistance the physician, Thomas Simon, devised a series of 30 tests, the chief purpose of which was to detect *subnormality*. After trying these tests on 203 school children in Paris, both Binet and Simon came to the conclusion that it was possible to devise a series of tests that would not only detect subnormality but also serve as a definite measure of mental unfoldment. With this thought in mind, they devised a scale consisting of 54 tests which they published in 1908. This scale was revised and republished in 1911.

Before taking up the problems confronting Binet and Simon in making an intelligence scale, a tabular synopsis of the 1911 Revision as adapted to American conditions will be presented for purposes of reference. A general description of the scale will then follow with an interpretation of a number of the descriptive terms used.

TABULAR SYNOPSIS OF THE BINET-SIMON SCALE, 1911 EDITION

Age 3:

1. Points to nose, eyes, and mouth.
2. Repeats two digits.
3. Enumerates objects in a picture.
4. Gives family name.
5. Repeats a sentence of six syllables.

Age 4:

1. Gives his sex.
2. Names key, knife, and penny.
3. Repeats three digits.
4. Compares two lines.

Age 5:

1. Compares two weights.
2. Copies a square.
3. Repeats a sentence of ten syllables.
4. Counts four pennies.
5. Unites the halves of a divided rectangle.

Age 6:

1. Distinguishes between morning and afternoon.
2. Defines familiar words in terms of use.
3. Copies a diamond.
4. Counts thirteen pennies.
5. Distinguishes pictures of ugly and pretty faces.

Age 7:

1. Shows right hand and left ear.
2. Describes a picture.
3. Executes three commissions, given simultaneously.
4. Counts the value of six sous, three of which are double.
5. Names four cardinal colors.

Age 8:

1. Compares two objects from memory.
2. Counts from 20 to 0.
3. Notes omissions from pictures.
4. Gives day and date.
5. Repeats five digits.

Age 9:

1. Gives change from twenty sous.
2. Defines familiar words in terms superior to use.
3. Recognizes all the pieces of money.
4. Names the months of the year, in order.
5. Answers easy "comprehension questions."

Age 10:

1. Arranges five blocks in order of weight.
2. Copies drawings from memory.
3. Criticizes absurd statements.
4. Answers difficult "comprehension questions."
5. Uses three given words in not more than two sentences.

Age 12:

1. Resists suggestion.
2. Composes one sentence containing three given words.
3. Names sixty words in three minutes.
4. Defines certain abstract words.
5. Discovers the sense of a disarranged sentence.

Age 15:

1. Repeats seven digits.
2. Finds three rhymes for a given word.
3. Repeats a sentence of twenty-six syllables.
4. Interprets pictures.
5. Interprets given facts.

Adult:

1. Solves the paper-cutting test.
2. Rearranges a triangle in imagination.
3. Gives differences between pairs of abstract terms.
4. Gives three differences between a president and a king.
5. Gives the main thought of a selection which he has heard read.

The Binet-Simon scale is an instrument for measuring *mental maturity*. By *maturity* we mean the development of *native capacity* as a whole by growth, training, and environment. *Mental growth* is a gradual increase of capacity for learning which comes as a result of the development of the nervous system apart from all training.⁹ By *native capacity* or *endowment* we mean the special capacity for functioning with which nature has provided the individual. Inborn capacity manifests itself only through learning. An individual born with a great capacity to become intelligent, but denied the opportunity to learn, would possess no intelligence. Intelligence must be acquired.

The tests in the Binet scale are arranged in progressive steps of increasing difficulty, each higher step involving

⁹ Leta S. Hollingworth, *The Psychology of Subnormal Children* (The Macmillan Co., 1920), p. 97.

the more tardily appearing functions, such as reasoning, complex comparisons, the associative functions, and the like, which depend directly on the maturation of the native capacities. The authors of the scale worked on the assumption that the intellectual ability of children of a given age tended to approach a relatively well-marked norm. The tests in each age-group were selected on this basis. The mental scale is merely the grouping together of individual tests in order to give a more general picture of the mental make-up of the individual. Binet originated the idea of grouping tests for estimating intelligence. For a long time he had been interested in the question of tests for various specific abilities. His work gradually led him to a study of individual cases, and, in summing up the psychological characteristics of individuals, as revealed by the mental tests, he came upon the idea of using a number of tests as a measure of the individual's capacity. In addition to this, his theoretical speculations as to what the tests were testing, led him to the conclusion that "attention" and "adaptation" were at bottom the chief factors that distinguished the intelligent from the unintelligent. The practical situation presented to Binet of separating the normal from the subnormal children of France called forth the first actual group of tests for differentiating intelligent and unintelligent children. He was called upon to discriminate between the normal and backward child, and the question was not whether this or that child was better in such a specific thing as memory or imagination, but whether the child was, in general, weaker in his intellectual endowment than the average child of his age. He therefore discarded the individual tests for specific ability and took a group of tests which seemed to cover in general the chief psychological characteristics that go to make up intelligence. It was Binet, therefore, who really *blazed the trail* through the jungle of mental measurements and

left us a path which leads to the general abilities of a child's mental life. As the norm or standard of intelligence, he took what the average child at each age could do.

These two points, *the use of the group tests and the average performance at each age as a standard of measurement, form the basic principles upon which all of our measuring scales of intelligence now rest.*

The Binet Tests Had Many Innovations.—Binet attempted to measure the higher thought processes instead of the simpler ones such as sensory discrimination, reaction time, retentiveness, and the like. He abandoned the older "faculty psychology" which had given direction to most of the testing up to this time and set problems for the reasoning powers—problems that provoke judgment about abstract matters and problems that draw on the discriminating powers of the examinee. If the faculties of the mind were separate and distinct so that they might be measured singly and then summated to get an idea of general intelligence, the problem might be simplified. But they are so interwoven and intertwined that they cannot be isolated for measurement. Memory, for instance, cannot be tested separate and apart from attention and other faculties because of the interfunctioning of these faculties. Of course, most mental phenomena elude absolute measurements in terms of amount; but they can be measured in terms of their relative magnitude. Although we cannot equate them, we can subject them to quantitative treatment.

The Constituent Functions of Intelligence Must Be Brought into Play.—Just as the total character of an individual cannot be determined by judging a single characteristic of his behavior, so his general intelligence cannot be determined by measuring the strength of one phase of his mental life. We must test many processes before a comprehensive idea of an individual's general intelligence

can be gained. For instance, if a completion test is used, would one be safe in saying that, because the child has the power to fill blanks in which words had been left out of a sentence or paragraph, he has a high degree of intelligence, whereas a child who cannot do this does not possess intelligence to the same degree? Or would a child's ability to repeat a number of words or figures be an infallible index of his general intelligence?

General intelligence is not simply the functioning of individual processes as such. It depends on their correlation and interfunctioning. While there was much criticism of the single tests and systems of tests used by the psychiatrist, which tested only one phase of the mind and attempted to judge the whole from this single characteristic, there is also much criticism of the Binet tests on the same general principles. This will be discussed more fully under the criticisms of the Binet scale.

Binet attempted to make a scale that would give a kind of composite picture of the individual's mentality by testing the strength of a limited number of these processes. The critics say that the scale is much better than the single tests of the psychiatrist but that it still fails to test enough characteristics to make the picture anything like complete. They say that because of the limited number of characteristics tested a lack of development of one trait, or an overdevelopment of another, tends to give a general result not in keeping with the facts.

The degree of interfunctioning between the various processes may be illustrated as follows: If a child's attention is *nil*, or nearly so, then his perceptive power does not focus long enough to produce the cortical alterations giving memory. Therefore, attempts at measuring memory also measure perception and other abilities of the mind. In this way a clue to the interfunctioning and interdependence of the various processes is obtained.

The Kind of Mental Functions Brought into Play.—Binet differed from the psychiatrist and other psychologists in the type of tests he designed. His tests were designed to test the higher and more complex thought processes such as reasoning power, abstract judgments, and the like, instead of attempting to measure sensory discrimination, the rapidity of reaction, and other lower and less complex powers. Up to this time it was considered impossible to measure these complex processes. The old "faculty psychology" had given direction to the earlier testing. It seemed to both the psychiatrist and the psychologist that sensory discrimination, memory, attention, and other processes of the mind, could be measured better if taken separately than if an attempt were made at summing the general results of all the different aspects of intelligence. This might be true were it not for the innerfunctioning of the various aspects of the mind which makes it impossible to completely isolate any one process for examination. Binet realized this fact and, instead of attempting to measure one aspect of the mind, he undertook to ascertain the *general level* of intelligence.

It is now generally conceded that such elementary processes as sensory discrimination (like distinguishing between two shades of color), reaction-time (as the rapidity with which an individual can tap), and visual acuity have but little to do with the higher thought processes, since many feeble-minded children have keen powers in sensory discrimination. It is apparently the power of comprehension, abstraction, and the ability to direct thought that separates the normal from the subnormal. Hence the other tests were very largely discarded by Binet. He sought to bring into play only those mental processes thought to be so closely concerned with *raw native ability* that they would give him an insight into this native endowment. If the percentages of passes did not increase in going from

younger children to older ones, the test was considered unfit since it did not indicate the degree of maturity of the developing intelligence. Or again, if children known to be bright passed a certain test more frequently than children known to be dull, such a test was considered satisfactory.

Establishment of the Zone of Normality.—Binet was confronted with a practical problem, that of separating the subnormal from the normal children in France. It is obvious that he could not detect the subnormal children unless he knew what was meant by normal ones. He had no precedent to follow in this difficult task. No one had determined what mental equipment a child must have to be considered normal. It was not sufficient to determine the mental equipment of a group of children of one age and take that as the standard, because the standard for each age had to be established if the degree of maturity of the child's native mental endowment was to be determined. It was therefore necessary to take a group of children for each age and determine normality for that age.

The idea of the age-grade method for measuring intelligence did not come to Binet until he had experimented with tests for fifteen years. The provisional scale published in 1905 did not employ the age-grade method but consisted of 32 tests roughly arranged in the order of their difficulty. Since no account is given as to how he came to employ the age-grade method, the supposition is that in working with the data gathered from the provisional scale made in 1905 he hit upon the age-grade idea. Suffice it to say that the age-grade idea was relatively complete in his 1908 scale.

Three problems confronted him in this task: (1) He must arbitrarily or otherwise choose for each age a group of children that he considered normal. (2) He must determine the general intelligence of each of these groups.

(3) He must determine the boundary lines which separate the *zone of normality* from the supernormal on the one hand and from the subnormal on the other.

In the solution of the first problem, if he were to choose the children from the higher classes in France his standard would be too high, assuming that the children from the upper classes had a larger stock of native mental endowment than those from the lower classes. On the other hand, if the children selected were known to be intellectually inferior, his scale would not be a fair measure of what a child's mental equipment really is. He apparently made his selection on the assumption that the normal child is the so-called average child; the common, ordinary child; the child who has that mental equipment possessed by the greatest number of children of that particular age. With the help of some of the teachers he then chose a group of children for each age which he thought represented the average child. His next problem was to determine the general intelligence of the groups chosen. This was done by a series of tests. He had tests of varying difficulty which he gave to each group. He found the number of children in each group who were able to pass the tests and thus, tentatively at least, determined the amount and degree of maturity of what he considered the native mental endowment of normal children. His method of determining what tests should belong to a particular age-level will be discussed later.

In order to further clarify the exact problems Binet had before him a diagrammatic representation of the range of mentality is given in Fig. I. In the great range of mentality from the idiot on the one hand to the genius on the other there seems to be a gradual increase of intelligence, and somewhere near the middle of this range is a section which may be designated as the *zone of normality*.

We may represent the range and distribution of mentality in Figure I thus: Let the left end of the line AB represent the lowest stages of mentality and the right end the highest. The curve is known as a *normal probability curve*, *normal frequency curve*, or *normal distribution curve*. It will be described more fully in a subsequent chapter. The number of cases having the various degrees of mentality is represented by the height of the curve above the base line AB . Thus, at the left the small number of people with extremely low mentality is represented by the

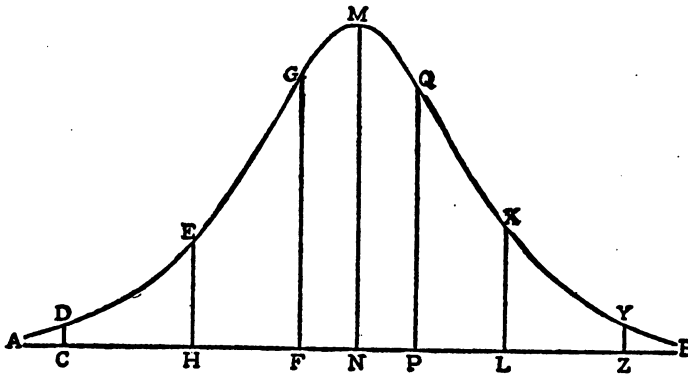


FIGURE I. ILLUSTRATING THE DISTRIBUTION OF MENTAL ABILITY ACCORDING TO THE NORMAL PROBABILITY CURVE

line CD . As the mental powers increase, the number of cases also increases until it reaches a maximum at point M . That is, a line drawn through point M perpendicular to the line AB is the longest line that can be drawn within the figure, and thus represents the degree of intelligence possessed by the largest group of people. Not only is it theoretically true that the distribution of intelligence is according to the normal frequency curve, but an actual count taken of any considerable number of unselected people shows that the distribution conforms remarkably well to the normal distribution curve. Thus the figures of

the Royal Commission of Great Britain, using the social-economic criterion of mental deficiency as a measure, show the distribution at the low end of the curve for certain districts surveyed to be: 585 idiots, 1,007 imbeciles, and 9,828 feeble-minded (morons)¹⁰ This, of course, is not an unselected group, but it does show the distribution among people whose mentality is below normal. The number of cases above normal then decreases until at the extreme right the curve comes down to the base line *AB*, or nearly so, because there are few extremely bright or intelligent people.

Near the middle of these two extremes along the line *AB* is a zone which we may call the *zone of normality*, the width of which may be represented by the line *FP*. The positions of the points *F* and *P* are arbitrarily chosen; hence the width of the zone of normality will vary according to the judgment of the scientist. It is important to note that normality does not mean a point on the scale but the distance between two points arbitrarily chosen. Reasonable latitude must be allowed for the pendulum of intelligence to swing a minor arc to the right or to the left and still remain within the zone of normality.

The third problem mentioned above, therefore, is settled in an arbitrary way. The zone of normality will vary in width according to the scientist, and no sharp line of demarcation will separate the normal from the subnormal.

Haberman defines a normal individual as "one whose reaction to given stimuli is no more or less in degree and manner than a certain quantum that we have become accustomed to."¹¹ This definition apparently conforms to Binet's conception of a normal individual.

¹⁰ The usual method of classifying people below the normal is to call those belonging to the lowest stage of mentality, idiots, and, in ascending order, imbeciles, feeble-minded, retarded, and normal.

¹¹ J. Victor Haberman, *The Intelligence Examination and Valuation*, p. 2.

Criteria for Separating the Normal from the Subnormal.—There are various ways of separating the normal from the subnormal. They are sometimes designated as: (1) *the social-economic criterion*; (2) *the pedagogical criterion*; (3) *the medical criterion*; and (4) *the psychological criterion*.¹²

The definition of a feeble-minded person given by the Royal Commission of Great Britain, which investigated the question of mental deficiency in 1904, illustrates the *social-economic* criterion for measuring intelligence. A feeble-minded person was defined as: "one who is capable of earning a living under favorable circumstances, but is incapable, from mental defect existing from birth, or from an early age, (a) of competing on equal terms with his normal fellows; or (b) of managing himself and his affairs with ordinary prudence."

The "jokers" in a definition of this kind are, of course, the expressions, "competing on equal terms" and "ordinary prudence." These allow a great deal of freedom in defining a feeble-minded individual.

The pedagogical criterion has been used for a long time as a basis for separating normal from subnormal children, the usual custom being to call children feeble-minded who are retarded pedagogically three years or more. This method is obviously defective because of the many things that might cause a child to be retarded three or more years and still have a normal mind or one almost normal. Sickness, a late start in school, bad eyesight, and poor economic conditions illustrate the point.

The medical criterion is based on the assumption that mental deficiency is analogous to physical disease. Binet has tersely stated the chief objections to the medical criterion as follows:

¹² Leta S. Hollingworth, *op. cit.*, p. 43.

Each one according to his own fancy fixes the boundary line separating these states. It is in regard to the facts that the doctors disagree. In looking closely, one can see that the confusion comes principally from the fault in the method of examination. When an alienist finds himself in the presence of a child of inferior intelligence, he does not examine him by bringing out each of the symptoms which the child manifests, and by interpreting all symptoms and classifying them; he contents himself with taking a subjective impression as a whole of his subject, and of making his diagnosis by instinct. We do not think we are going to far in saying that at the present time very few physicians would be able to cite with absolute precision the objective and invariable sign or signs by which they distinguish the degrees of inferior mentality.

The physician has been trained to diagnose and treat physical disorders primarily, and when confronted with a mental condition reasons very largely from analogy of what he knows of physical states.¹³

It is, of course, the *psychological criterion* in which we are primarily interested. Each of the other attempts at classifying individuals has depended so much upon the judgment of those making the classification that it is far less definite than it should be for practical use.

A knowledge of the distribution of intelligence among the people of a nation is of great importance both theoretically and practically. It is worth a great deal to know what percentage of the people may fall within the range of normality and to find out what positions on the scale for testing mental development are symptomatic of social deficiency. It is also important, from a sociological standpoint, to know that adults testing below a certain score are so low in intellectual development that it is a question as to whether they have sufficient equipment to survive socially. Adults who test only ten years old mentally, for instance, are an uncertain group in intellectual

¹³ *Ibid.*, pp. 47-48.

ability with the probability that they will require more or less social care, while those who test only nine years old are deficient enough to need continuous care. Goddard found no case at the Vineland School for the Feeble-Minded which tested higher than 12 mentally. Huey found but two such cases in the State Asylum at Lincoln, Ill., and Kuhlmann found only ten at the Minnesota State School for the Feeble-Minded.¹⁴

In the distribution of intelligence Terman found that about 60 per cent of all school children test between 90 and 110 I. Q. ("I. Q." means *intelligence quotient* and is the quotient found by multiplying the mental age by 100 and dividing by the chronological age). About 40 per cent test between 95 and 105. An intelligence quotient of 110 to 120 is five times as frequent among children of superior social status as among those of inferior social standing; the proportion among the superior social group being 24 per cent of all, whereas but 5 per cent of those that belong to the inferior social group test with an I. Q. of from 110-120. Not more than three children out of 100 score as high as 125 I. Q. and only one in 100 as high as 130, while an I. Q. of 140 is made by only one in 250 to 300.¹⁵ Terman makes the following classification of intelligence quotients (from the Stanford Revision):¹⁶

I. Q.	Classification
Above 140.....	"Near" genius or genius.
120-140	Very superior intelligence.
110-120	Superior intelligence.
90-110	Normal, or average, intelligence.
80- 90	Dullness, rarely classifiable as feeble-mindedness.
70- 80	Borderline deficiency, sometimes classifiable as dullness, often as feeble-mindedness.
Below 70	Definite feeble-mindedness.

¹⁴ Cf. James B. Minor, *Deficiency and Delinquency*, p. 95.

¹⁵ Terman, *op. cit.*, pp. 94-96.

¹⁶ *Ibid.*, p. 79.

Feeble-minded individuals with intelligence quotients between 50 and 70 include most of the morons; those between 20 and 50, imbeciles; and those below 20, idiots.

Are Differences in Intelligence of Degree or of Kind?—

It makes a great deal of difference to those attempting to measure intelligence whether the differences found among individuals as to their intelligence is a difference in *degree* or a difference in the *kind* of mental traits and characteristics individuals have.

If the difference is one of kind then we might expect the normal individual to have certain mental traits, powers, and characteristics not possessed by the subnormal individual. On the other hand, if it is simply a difference in degree, then the idiot possesses some of all the mental traits, powers, and characteristics that the normal individual or even the genius has. And such would now seem to be the consensus of opinion, for a search for a qualitative difference between feeble-minded and normal individuals has failed to disclose any characteristic or mental trait found in the normal individual and not possessed by the feeble-minded to at least a small degree.

In his early work even Binet considered the difference between normal individuals and subnormal ones to be *one of kind*. In his work entitled *Mentally Defective Children* he says:

A second and totally different theory is tenable, and this one appears to us to be much nearer the truth. It is that a defective child does not resemble in any way a normal one whose development has been retarded or arrested. He is inferior not in degree, but in kind. . . . An unequal and imperfect development is his special characteristic. These inequalities of development may vary to any degree in different subjects. They always produce a want of equilibrium, and this want is the differentiating attribute of the defective child.

In Binet's later works he came to the conclusion that the difference between the normal and subnormal was not

one of kind but one of degree. In his volume entitled *The Intelligence of the Feeble-Minded* he writes: "We may thus pass in review all our faculties, and determine that not one is entirely lacking in them. . . . They always have them in some degree. . . . The arsenal of their intellect is equipped with all the weapons."

Dr. Norsworthy published the results of her experiments in measuring feeble-minded children in 1906. She tested a variety of their mental traits, and also the same traits of a number of normal school children. In no case did she find the normal child having mental traits not possessed by the feeble-minded child. The experiments of Pearson and Jaedorholm coincide with the findings of Norsworthy and also the later conclusions of Binet. The conclusions of these scientists are further confirmed by statistics. The subnormal children occupy the lower end of the normal distribution curve and the greater the degree of subnormality the fewer the cases found. If they were in a class by themselves, the number of cases would be less near the limits of the class and greatest near the middle. But, as was stated above, the number grows progressively greater from the lowest mentality to the normal individual.

Choosing Tests to Measure Intelligence.—The selection of tests that will measure general intelligence is an extremely difficult problem because of the complex and subtle character of the mind to be measured. Many psychological principles must be kept in mind. Just any kind of tests will by no means satisfy the conditions. We indicated earlier in the chapter that the type of test chosen depended on what was conceived to be the nature of intelligence. In the earlier testing work, for instance, Binet believed that the essence of intelligence was the capacity to adjust the attention. He therefore devised tests such as the cancellation of letters in a specially prepared sheet, so as to bring into play this faculty. He thought that the ability to dis-

criminate between two near-lying points of the compass on the skin was a matter of attention rather than sensation and used this as a test in the earlier part of his work. We shall enumerate some of the problems to be solved in the selection of tests and criticize briefly their attempted solutions.

Tests Must Not Be Influenced by External and Chance Conditions.—If the tests were affected to any great degree by school training or environmental conditions they would not be a measure of native endowment. Measuring general intelligence is not merely measuring what a child knows or has retained; it is not a matter of knowledge but of something essentially dynamic behind and veritably in this knowledge. It is interfunctioning, harmonious and correlated in the normal, but discordant and disrelated in the psychopathic, the intellectually defective.

Ayres¹⁷ criticizes the Binet tests on the ground that five of them depend upon the child's recent environmental experience. It may be, however, that his criticisms are not just because the degree to which environment is able to affect a child may reveal his native endowment. Environmental effects are determined and distinguished in various ways from innate capacity.¹⁸ One method is the effect of practice on performance. Other things being equal, the less practice required to make a given unit of gain the greater the initial endowment. Great superiority over the average in intellectual effort indicates a high degree of native endowment. Also the appearance of definite activities, such as when a child spontaneously interests himself in music or literature, indicates that the child possesses strong innate tendencies in those directions. When practice in varying amounts has already influenced initial

¹⁷ *The Binet-Simon Measuring Scale of Intelligence; Some Criticisms and Suggestions.*

¹⁸ Cf. Meumann, *Vorlesungen*, Vol. II., pp. 306-314.

capacities, the cause of the differences in individuals can be inferred from the effects of equalizing practice.¹⁹ Further practice of equal amounts will reduce differences where these are not due to innate causes; if, however, the inequalities increase with further practice, they are not caused by irregularities in previous practice and are therefore due to innate conditions.²⁰

Only Those Tests Must Be Chosen That Afford a Decided and Reliably Symptomatic Value, General Applicability and Possibility of Objective Evolution.—

As a matter of fact, a scale for the measurement of intelligence is more limited in scope than the above description would suggest, since it omits a great many capacities or abilities that are not supposed to be indicative of the mentality of an individual. For example, there are tests for the ability to discriminate two points on the skin and for the ability to discriminate different shades of color; but we do not include these tests in our scales of intelligence, because it is not believed at the present time that such tests have diagnostic value for distinguishing between different grades of intelligence.

Tests Must Not Depend Too Much on the Ability to Use Language.—Just how much the ability to use language is indicative of intelligence is a question. Have we a valid test when the ability to pass it depends not merely on the comprehension of language but also upon the ability to frame the adequate language response? If one depends on definitions too exclusively, he may be measuring school achievements rather than raw intelligence. It is evident that it would limit and perhaps vitiate the results a great deal if too much emphasis were put on the language factor.

¹⁹ Edward L. Thorndike, *Educational Psychology*, Vol. III, p. 205.

²⁰ Robert R. Rusk, *Experimental Education* (Longmans, Green & Co., 1919), p. 162.

This language difficulty inherent in the Binet Scale and in all the revisions of it became very pronounced as soon as the use of the scale spread to workers in the various fields. The clinical psychologists in the large cities were face to face with the problem of the foreign child, the speech-defective, and the deaf children. It was obvious that the Binet Scale was not adequate for the mental examination of such cases. Other tests not involving language were introduced, and these gave rise to the type now generally known as *performance tests*, which will be described later.

The essential characteristic of this type of test is that it shall not require any kind of language response on the part of the child for an adequate performance of the test. It is true that seven of the Binet tests depend on a child's ability to read and write, and many others require him to express himself through language. The degree to which this scale is vitiated because of its dependence on language has not been determined. It has been determined, however, that there is a high correlation between the ability to use language and general intelligence.

Determining the Age to Which a Test Should Be Assigned.—One of the difficult problems in devising a mental scale has been the determination of the particular age to which the parts should be assigned. Suppose, for instance, we take one of the questions Binet assigned to the five-year-old group. The examiner repeats a sentence with ten syllables in it and asks the examinee to repeat it after him. Suppose 15 per cent of the children three years old can do it successfully, 40 per cent of the four-year-old children, 75 per cent of the five-year-olds and 100 per cent of the six-year-olds. To what grade should this sentence be assigned? Shall we assign it to the six-year-old group because this was the lowest group that was able to make a score of 100 per cent? Or, if 75 per cent of them are

able to do it, is that sufficient? Or, if any percentage less than 100 is used, what percentage should it be before it is assigned for that particular age?

In the standardization of individual tests the usual custom has been to consider a test properly placed if it is passed by 75 per cent of the children. The idea is that if the test is properly placed it will be passed by 50 per cent of the children who may be considered of average intelligence, plus 25 per cent of the children who are considered above the average.²¹ Of course, theoretically at least, 25 per cent who are assumed to be below normal will be unable to pass the test.

Binet's guiding principles in the arrangement of tests were: (1) Find an arrangement of the tests which would cause the average child of any given age to test "at age"; that is the average six-year-old must show a mental age of six years, the average eight-year-old a mental age of eight, and so on. (2) In order to obtain this result he found that it was necessary to locate an individual test in that year where it was passed by about two-thirds to three-fourths of the unselected children.

Terman²² considers the proper assembling of the tests one of Binet's biggest problems and one in which he failed in many instances, since many of the tests were misplaced as much as one year and one was misplaced six years.

There are many objections to this method of locating tests; but on the whole it seems to be about as good as we are able to get at the present time.

Problems in Scoring.—There are certain problems incident to scoring the tests which deserve special treatment. We shall state briefly what these problems are and note Binet's solution of them.

²¹ S. D. Porteus, *Condensed Guide to the Binet Tests* (published by the Training School, Vineland, N. J., April, 1920), Part I, p. 10.

²² Terman, *op. cit.*, p. 48.

The All-or-None Method in Scoring.—The question here is as to whether or not credit should be given for a test done only in part. For instance: A child is asked to count backwards from 20 to 0. Suppose he makes three errors but counts the rest correctly. Should any credit be given for such work? Binet said no. He held that a child must complete each test before any credit is given. This method of scoring has been severely criticized by many writers, notably Yerkes, Bridges, and Hardwick.²³ The merits of the all-or-none method of scoring will be discussed more fully under *A Point Scale for Measuring Mental Ability* in the next chapter.

Shall a Child Be Required to Pass All the Tests at Each Age-Level?—As was indicated above, each age-level had from five to seven tests. The question arises as to whether or not it is necessary for a child to pass all the tests at each age-level before he shall be allowed to try those of a higher age-level. Binet realized that children had lapses in attention and that all the processes of the mind did not develop with the same rapidity. Therefore, in order to take cognizance of these characteristics, he arbitrarily said that if a child passes all the tests in any age-level save one, he shall be considered to belong to that age-level. For instance, at the five-year level there are five tests. A child passing four of them would be considered five years old mentally. But suppose a child should pass all the tests in the five-year level, and two in the six-year level, how would we reckon his age? Binet solved this problem as follows: He took as a basis the highest age at which the child made a perfect score; that is, the age at which he passed all the tests or all save one. Then for every five tests he passed beyond this age, one more year was added to his mental age.

²³ *A Point Scale for Measuring Mental Ability* (Warwick & York, 1915).

It sometimes happens that a child will pass all the tests in the five-year group, for instance, fail in two or three of them in the six-year group, and then pass all in the seven-year group. In a case of this kind the problem is, What age-level shall be taken as the basis for computing mental age? Yerkes and Bridges especially call attention to this point.

With What Tests Shall the Examination Begin?—Suppose one were called upon to test a boy eight years old, where on the scale should he begin? Should he begin with the easiest tests, that is, the tests for the three-year-olds, and go as high on the scale as the boy can go, or should one begin with some other age-level, as the eight-year-old level, and determine the boy's mental age by going up if the boy can pass tests above that age, or down if he is unable to pass the tests at this age-level? Binet solved this problem by the trial-and-error method. Taking one case with another, his best guess would be that the child is normal or nearly so. Therefore he would start enough below the normal age to find an age group he was pretty sure the child could pass and proceed with the tests until the child was unable to pass any tests of a higher age-level.

In making these tests it was assumed that a certain mental level normally goes with a certain chronological age, so that the relation of mental age to chronological age indicates the amount of discrepancy between the amount of intelligence present and that required for normality. The custom has been to compute, in a simple way, the difference between the two ages, which, when negative, gave the absolute mental retardation and, when positive, the mental acceleration. This method is open to criticism, however. It has been shown that the increments from age to age are not the same and that a child chronologically ten years old and retarded two years does not have the same retarda-

tion as a child chronologically eight years old and retarded two years.

The "At Age" and the Normal Child.—Care must be taken to distinguish between the "*at age*" child and the *normal* child. The former is one whose mental and chronological ages are the same. But a child may be considered normal if he is mentally a few months younger or older than the "*at age*" child. In other words, the "*at age*" children simply constitute the middle section of the *zone of normality* although children advanced or retarded a few months are still within the accepted bounds of normality.

Binet Tests Give More Than a Composite Picture.—While it is the purpose of the Binet Scale to give a composite picture of a child's mentality, it nevertheless gives valuable insight into some of the detailed aspects of his mental functioning. Meumann points out that there are three distinct types of tests: (a) tests of capacity or endowment, (b) tests of maturity or development, and (c) tests of environment or training. Bearing in mind this analysis of the scale the examiner may learn not only that the child is either normal, subnormal, or supernormal, but he may also learn in the case of subnormality, for instance, whether the defect is due to training or to native capacity. The giving of a conventional list of facts displays the *quality of training*, for instance, whereas the repetition of auditory digits or sentences displays, inferentially, at least, native capacity.

Coefficient of Mental Age and the Intelligence Quotient.—Binet's method of expressing the relation between a child's mental and chronological ages was simply to say that the child was six years old chronologically, for instance, and six-and-a-half years old mentally, or six years old chronologically and seven-and-a-half years old mentally, as the case might be. Stern suggested a new method

of evaluating the results of the scale. His plan was to divide the number of tests a child actually passed by the number he ought to pass and call the quotient the *coefficient of mental age*. By this scheme a child six years old chronologically, for instance, should pass all the tests up to and including those of the sixth-year group, or 20 in all. If he passed but 15, his coefficient of mental age would be 0.75. If, however, he should be seven years old mentally, his coefficient of mental age would be 1.17, etc.

Dr. Terman uses the *intelligence quotient*, sometimes spoken of as the I. Q., for the purpose of expressing the relation between the chronological and mental ages of children. It is found by dividing the mental age by the chronological age just as Stern suggested in determining the coefficient of mental age; but instead of using fractions to represent this relation, Terman multiplied the quotient by 100 to avoid them. Thus the "at age" child by the Terman method of recording age is given an I. Q. of 100, and any child is considered normal whose I. Q. is between 90 and 110.

Limitations of the Tests.—The Binet Scale does not pretend to measure the entire mentality of the subject. The emotions, for instance, are measured only in the remotest way, if at all. It is not claimed that the scale reveals special talent, and for this reason will not serve as a detailed chart for the vocational guidance of children. *It will, however, roughly bound the limits within which one's intelligence will permit success.* Sharp lines of demarcation are not to be expected between the various degrees of mentality.

The Age of Mental Maturity.—Like other body tissues, the nervous system, which is the physiological basis of mental life, does not grow indefinitely. Since psychologists cannot directly measure the course of growth of the nervous system in living children, they measure it indirectly by

measuring the *behavior* of the child in taking mental tests. Scientists are not agreed, from tests thus far made, as to just when "native intelligence," or mental growth, has reached maturity. Practically all are agreed simply that maturity comes somewhere in the 'teens and that beyond this there is little or no development of this native capacity. Terman says:²⁴

Native intelligence, in so far as it can be measured by tests now available, appears to improve but little after the age of 15 or 16 years. It follows that in calculating the I. Q. (intelligence quotient) of an adult subject it will be necessary to disregard the years he has lived beyond the point where intelligence attains its final development. Although the location of this point is not exactly known, it will be sufficiently accurate for our purpose to assume its location at 16 years. Accordingly, any person over 16 years of age, however old, is for purposes of calculating I. Q. considered to be just 16 years old. If a youth of 18 and a man of 60 years both have a mental age of 12 years the I. Q. in each case is $12 \div 16$ or 0.75.

Porteus claims that the recent work of the army examiners has shown that the age of maturity is much below the level set by Terman and that if the Terman standards are used, it is possible to diagnose an adolescent as being at the feeble-minded level when as a matter of fact he is comparatively little below the average of the ordinary population.

Yerkes and Bridges say in connection with this question that "it seems highly probable that the adult level is attained as early as the sixteenth year."²⁵

Spearman and Kuhlmann think that at the age of 15 the native intelligence is mature. The former says the fact that mental ability reaches its full development about the period of puberty is still further evidenced by

²⁴ Terman, *op. cit.*, pp. 140-141.

²⁵ *A Point Scale for Measuring Mental Ability*, p. 64.

physiology. For the human brain has been shown to attain its maximum weight between the ages of 10 and 15 years.²⁶ On the other hand, Wallin thinks we should have more evidence before choosing a fixed age for the maturity of native intelligence.²⁷

Criticisms of the Binet Tests.—The criticisms of the Binet tests are many and varied. Some are favorable and some are unfavorable. We shall note first some of the unfavorable criticisms. One of the most radical and severe criticisms of the Binet Scale has been made by Haberman, who, speaking of the Binet tests, says:²⁸ "Like the Freudian delusion, this Binet dilettanteism has taken us by storm and shows splendidly our native gullibility and likewise an interesting phase of the hysterical lay-medical activity of the times."

One of the early critics of the Binet Scale was Dr. Leonard P. Ayres. He criticized the tests from the standpoint of their content. His criticisms fall under six heads: (1) The tests depend predominantly on the child's ability to use words fluently, and only to a limited extent to perform acts. (2) Five of the tests depend upon the child's recent environmental experience; hence it is questionable whether they test native endowment. (3) Seven of them depend upon his ability to read and write, which again raises the question as to whether they test native ability or school achievement. (4) Too great weight is given to tests of ability to repeat words and numbers. (5) Too great weight is given to "puzzle tests." (6) Unreasonable emphasis is given to tests of ability to define abstract terms. Since Ayres made his criticisms in

²⁶ C. Spearman, "The Heredity of Ability," *Eugenics Review*, 1914, pp. 229-237.

²⁷ J. E. Wallace Wallin, "Re-Averments Respecting Psychoclinical Norms and Scales of Development," *Psychological Clinic*, 1913, pp. 89-97.

²⁸ Haberman, *op. cit.*, p. 14.

1911 much has been learned about the Binet tests, and we now know that many of the criticisms made by him are not so serious as they looked at that stage of their development.

The main defects of the Binet Scale according to Meumann are: (1) The single tests are not rightly graded according to their difficulty. (2) The tests of each kind are not sufficiently numerous, and those of a particular kind are not repeated year after year so as to trace the child's development in the various faculties of the mind. The different age-groups deal with entirely different mental functions; for instance, one of the tests in the five-year group is to repeat a sentence of ten syllables. This is a test in auditory memory. The next time the child is called upon to show his ability in auditory memory is when he is asked in the eight-year group to repeat five digits. The interval is three years and cannot, therefore, be a very accurate measure of a child's rote memory development. (3) The tests determine quite different capabilities; that is, they are not systematically arranged according to definite points of view. (4) The exact sum of the whole testing has not been decided upon.

The scale is further criticized on the ground that there are two tests in some age-groups which depend on memory. This gives undue emphasis to one particular trait to the exclusion of others, which may be as much of a measure of mentality as memory.

Limits of Traits Not Determined by the Binet Scale.—

The Binet Scale fails to determine the limits of a trait for two reasons: (1) The unit of accomplishment is so large that small degrees of progress are not recorded. (2) In one age-group auditory memory, for instance, may be tested whereas in the next group a different kind of memory is tested. If the child passes the test in auditory memory he is not tested again in that particular trait until three

or four age-groups are passed. Hence his ability in that particular trait is not determined by the test taken.

Myers²⁹ points out that the determination of endowment and the measure of general intelligence has been approached in a more strictly psychological manner by tests different from the Binet tests. In his opinion, the Binet tests are *tests of production* rather than psychological tests. He says: "They determine how *much* an individual can work, how *much* he knows—not *how* he works, *how* he knows. A man's productivity, of course, is what we want to ascertain in everyday life. We do not care how a man comes to use or acquire his powers; we are content with a mere dynamometric or other record of his prowess. But this aspect cannot properly be called the psychological aspect."

The Binet Scale Criticized on Other Points.—The scale is further criticized because there is not the same number of tests at each age-level. The four-year-old group contains but four tests whereas the other groups contain five; and it may be easier to pass three tests out of four than four out of five.

The absolute amount of retardation, that is, the difference between the mental and chronological ages, is not a comparable quantity when different age levels are in question. Or again, two children with the same mental and chronological ages may have a very wide variation in the degrees of maturation of their native capacities. For instance, a boy ten years old mentally and chronologically may have passed all the tests consecutively up to and including the ten-year group but be unable to go beyond that age-level. Another boy, mentally and chronologically the same age as reckoned by the Binet Scale, may have been unable to pass all the tests in each

²⁹ *British Medical Journal*, No. 2613, pp. 196-197.

age group beyond the eighth year, but because he was able to pass ten tests above the eight-year level he was given a mental age of ten the same as the other boy. The range of the distribution of passes and failures is, as a rule, very much wider in the feeble-minded than in the normal individual. That is, a subnormal child may be very proficient in one trait and deficient in another. Some investigators have found it to be twice that of the normal individual.

A mind with the least variation from the normal seems to be of a higher type than one which varies a great deal; for example, a child who passes all the tests in the eight-year group and one in the nine-year group and can go no further would be considered of a higher type than a child who passes all the tests in the seven-year group, two in the eighth, and one in each of the next three groups, making him eight years old mentally.

Many criticize the Binet Scale because it does not display more completely the character of the mind being tested. Seashore, Pyle, and others demand a more fundamental analysis and a more exact determination of mental ability than is possible to obtain by the Binet tests. Thus Seashore writes:³⁰

Retardation does not follow a common flat level any more than growth does, nor even nearly so much. A child develops one capacity several times as fast, and often at the expense of another faculty. This differentiation is even more striking in retardation. What is more, those who employ the tests for practical purposes should not be satisfied with a flat mental age. . . . In a study of the normal individual we seek to discover *fortes* and his faults, in short, to discover his particular deviation from the norm of the common level. There is no reason why the Binet-Simon tests should not develop into specific measures of the relative rank, or age, of more specific capacities and powers, such as reasoning ability, sensory observation, memory, imagina-

³⁰ *Journal of Educational Psychology*, Vol. 3, p. 50.

tion, initiative, emotional life, self-control, etc. A child may be at the mental age of six in one capacity and twelve in another, and the important thing to know about the individual is this difference and direction of unsymmetrical development. It may be that a general flat-age test must be retained for certain purposes, but even that must be interpreted in the light of measures of specific capacities. Only by extension in recognition of this principle can any set of tests be of permanent value.

Pyle criticizes the tests on the ground that there is no common plan running through the tests for the successive years. He argues for "a series of tests for determining the degree of development of logical memory, rote memory, attention, imagination, association, and two aspects of mind more complex, learning capacity and reasoning. . . . It is more important, it seems to me, to know specifically the condition of the child with reference to the development of the separate mental traits than to know his average performance with respect to them all."³¹

Porteus³² has pointed out some of the limitations of the Binet tests as follows: They do not constitute a perfect instrument of research and have failed to fulfill many expectations founded on them. They cannot be relied upon for accurate diagnosis of the highest grades of feeble-mindedness; he points out that there are other factors besides the degree of intellectual development which have a bearing on social competency. Many intellectually inferior persons are judged normal by social criteria because they possess practical abilities not evaluated by Binet. On the other hand, there are intellectually normal individuals who are socially unfit because of instability, weakness of temperament, volition, and other traits not revealed by the Binet examination.

³¹ *Ibid.*, Vol. 3, pp. 95-96.

³² *Condensed Guide to the Binet Tests*, Bulletin No. 19, Training School at Vineland, N. J., pp. 1-3.

He further criticizes the tests on the ground that they are too literary. Fifty-seven questions out of 74 require an oral language response. Language development, either from the standpoint of comprehension, range of vocabulary, description, or defining power, is the main capacity tested in 50 per cent of the cases. Thirty per cent of the tests depend on previous educational training. Forty-eight per cent of the tests depend on mediate and immediate memory.

His general criticisms are that the tests are too literary; that they favor the glib-tongued, quick-thinking child who has had good educational advantages, who memorizes easily and therefore shows good scholastic promise. This is why the Binet tests correlate so highly with school training. It has been mainly through comparison with teachers' judgments that the validity of the tests has been established. There are cases, however, in which children with high Binet records show little power of adaptation to school conditions and also cases in which the class dullard "makes good." The main point that Porteus is pointing out is that the Binet tests will not measure accurately all individuals. On the other hand, it is not necessary to stretch out one series of tests so thin as to cover the whole field anyway. The limitations of the tests pointed out by Porteus does not mean that the Binet tests are not to be used in diagnosis but that they must not be used to the exclusion of all others for diagnostic work.

These defects were not unknown to Binet. In regard to mental age he writes thus:³³ "It has no bearing on the cause of retardation, nor upon its peculiar nature, nor upon the means of rectifying it." In regard to the fallacy of regarding retardation as merely equivalent to a lower mental age, he says:³⁴

³³ *Mentally Defective Children*, English trans., p. 12. ³⁴ *Ibid.*, p. 13.

A defective child does not resemble in any way a normal one whose development has been retarded or arrested. He is inferior not in degree but in kind. The retardation of his development has not been uniform. Obstructed in one direction, his development has progressed in others. To some extent he has cultivated substitutes for what is lacking. Consequently, such a child is not strictly comparable to a normal child younger than himself.

Some Favorable Criticisms of the Binet Scale.—In spite of its defects the Binet Scale has been applied in practically every civilized country and has received general approval. Its shortcomings are in its details and not in the fundamental principles. The opinions of those best qualified to sit in judgment on it are to the effect that it has demonstrated its value and that it is destined through its revisions and corrections to be a most valuable instrument in measuring of mentality.

Kuhlmann writes thus:³⁵

There can be no question about the fact that the Binet-Simon theses do not make half so frequent or half as great errors in the mental ages [of feeble-minded children] as are included in gradings based on careful, prolonged general observation by experienced observers.

Meumann writes as follows:³⁶

All the different authors who have made these researches [with Binet's method] are in a general way unanimous in recognizing that the principle of the scale is extremely fortunate, and all believe that it offers the basis of a most useful method for the examination of intelligence.

Stern says:³⁷

That, despite the differences in race and language, despite the divergences in school organization and in methods of instruction,

³⁵ Dr. F. Kuhlmann, "The Binet and Simon Tests of Intelligence in Grading Feeble-Minded Children," *Journal of Psycho-Asthenics*, 1912, p. 189.

³⁶ Ernest Meumann, *Experimentelle Pädagogik* (1913), Vol. II, p. 277.

³⁷ Stern, *op. cit.*, p. 49.

there should be so decided agreement in the reactions of children—is, in my opinion, the best vindication of the *principle* of the tests that one could imagine, because this agreement demonstrates that *the tests do actually reach and discover the general developmental conditions of intelligence* (so far as these are operative in public-school children of the present cultural epoch), and not mere fragments of knowledge and attainments acquired by chance.

Goddard says:³⁸

It is without doubt the most satisfactory and accurate method of determining a child's intelligence that we have, and so far superior to everything else which has been proposed that as yet there is nothing else to be considered.

Goddard not only defends the Binet Scale as a whole but defends the age grouping on the ground that it conforms to the normal distribution curve. He says that if the questions were not properly grouped, age for age, but were too hard or too easy, the largest group would not be one "at age" but would be a year below or a year above according as they were too hard or too easy.

Other Problems That Confront Those Testing Intelligence.—A number of other interesting and perplexing problems confront those attempting to measure the growth of general educational capacity.

1. *Does the defective progress normally to a certain point and then suffer arrest, or has mental growth been retarded from birth?*—The study of subnormal children shows clearly that they are inferior and below par from the beginning. They suffer no arrest in their development any more than a normal child does. They simply develop more slowly and hence never reach the stage normal children reach when they are mature. This applies not

³⁸ H. H. Goddard, "The Binet Measuring Scale of Intelligence. What It Is and How It Is to Be Used," *Training School Bulletin*, Vineland, N. J., 1912.

only to their mental development but also to their physical development. They are slower in learning to walk, talk, creep, sit up, the appearance of their teeth, etc.

2. *Does the defective child have the same mental equipment as the normal child of the same mental age?*—According to the best evidence that can be secured along this line, the defective child seems to have a mental equipment different from a normal child of the same mental age. For example, a child twelve years old chronologically and seven years old mentally differs from a normal seven-year-old child in certain specific habits and bits of information. The defective child has accumulated certain of these habits and bits of knowledge simply because he has lived longer and has had more experience. His native endowment is more nearly matured than that of the normal child. But maturity in the defective child carries him only part way up the scale and it is in this sense that he is mentally equal to the normal child. The instincts differ, especially when a defective adolescent is compared with a normal child of the same mental age.

3. *Do feeble-minded children mature mentally at the same chronological age as normal children?*—Measurements taken in training schools show that defective children continue to grow mentally until well advanced in adolescence. Their development is slow, and there is relatively more variation in the development of the traits than in normal children, but it may be said to be continuous.

4. *Are subnormal children equally deficient in all abilities?*—Just as normal children develop one ability more than another, so subnormal children may be quite proficient in some intellectual traits and very deficient in others. Visual acuity and certain types of memory, for example, may be as well developed in the subnormal child as in the normal and in some cases even better. Indeed,

it is this marked variation of mental abilities from the norm that is in part responsible for subnormality. Not only do the abilities, taken as a class, develop slowly but very irregularly. His several intellectual abilities reach several levels on the intellectual scale.

Perhaps enough has been said about the Binet Scale and the many problems that confront those attempting to use it. We have given a brief history of some of the more important problems and criticisms of the Binet Scale. All are agreed that it is far from a perfect measuring instrument. It has been open to many criticisms. Yet in spite of the criticisms, the scale, nevertheless, enables one to make a rough classification of pupils in a comparatively short time and by rather simple means. Its value is perhaps practical and pedagogical rather than theoretical and psychological. There is no question about its being considerably in advance of the estimates of intelligence based on school performance or on the biased judgments of teachers or parents.

A summary criticism and an evaluation of attempts to measure intelligence will be given at the end of the next chapter.

BIBLIOGRAPHY

1. AYRES, LEONARD P., *The Binet-Simon Measuring Scale for Intelligence: Some Criticisms and Suggestions* (Russell Sage Foundation, New York, 1911).
2. BINET, ALFRED, "L'intelligence des imbéciles," *L'Année Psychologique*, 1909.
3. BURT, CYRIL, "The Measurement of Intelligence by the Binet Tests," *Eugenics Review*, 1914, pp. 6, 36-50, 140-152.
4. GODDARD, H. H., "The Binet Measuring Scale of Intelligence; What It Is and How It Is to Be Used," *Training School Bulletin*, Vineland, N. J., 1912.
5. HOLLINGWORTH, LETA S., *The Psychology of Subnormal Children* (The Macmillan Co., 1920).

THE MEASUREMENT OF INTELLIGENCE 103

6. HABERMAN, J. VICTOR, *The Intelligence Examination and Evaluation* (American Medical Association, Chicago, 1915), Pamphlet, 16 pages.
7. KUHLMANN, F., "The Binet and Simon Tests of Intelligence in Grading Feeble-Minded Children," *Journal of Psycho-As-thenics*, 1912, Vol. 16, pp. 173-193.
8. MINOR, JAMES BURT, *Deficiency and Delinquency* (Warwick & York, 1918).
9. PORTEUS, S. D., *Condensed Guide to the Binet Tests*, Part I (Training School, Vineland, New Jersey, 1920).
10. PYLE, W. H., "A Suggestion for the Improvement and Extension of Mental Tests," *Journal of Educational Psychology*, Vol. 3, pp. 95-96.
11. RUSK, ROBERT R., *Experimental Education* (Longmans, Green & Co., 1919).
12. SPEARMAN, C., "The Heredity of Ability," *Eugenics Re-view*, 1914, pp. 219-237.
13. SEASHORE, C. E., "The Binet-Simon Tests," *Journal of Educational Psychology*, Vol. 3, p. 50.
14. SCHWEGLER, RAYMOND A., *The Binet-Simon Scale of In-telligence* (University of Kansas, 1914).
15. STERN, WILLIAM, *The Psychological Methods of Testing Intelligence*, translated by Guy Montrose Whipple (Warwick & York, 1914).
16. TERMAN, LEWIS M., *The Measurement of Intelligence* (Houghton Mifflin Co., 1916).
17. THORNDIKE, EDWARD L., *Educational Psychology*, Vol. 3 (Teachers College, Columbia University, 1914).
18. TITCHENER, E. B., "Wilhelm Wundt," *American Journal of Psychology*, Vol. 32, April, 1921, pp. 161-178.
19. WALLIN, J. E. WALLACE, "Re-Averments Respecting Psycho-clinical Norms and Scales of Development," *Psychological Clinic*, Vol. 7, 1913, pp. 89-96.
20. WHIPPLE, GUY MONTROSE, *Manual of Mental and Physical Tests*, Parts I and II (Warwick & York, 1915).
21. YERKES, BRIDGES, and HARDWICK, *A Point Scale for Measuring Mental Ability* (Warwick & York, 1915).

CHAPTER IV

THE MEASUREMENT OF INTELLIGENCE—*Continued*

The discussion of the measurement of intelligence in this chapter will be divided into three parts: (1) the extension and revision of the Binet Scale and other measures of intelligence; (2) group intelligence scales; and (3) a summary and evaluation of measures of intelligence.

I. THE EXTENSION AND REVISION OF THE BINET SCALE AND OTHER MEASURES OF INTELLIGENCE

Among the many revisions and extensions of the Binet Scale, perhaps the most important one in America is that made by Dr. Lewis M. Terman, known as the *Stanford Revision*. Other noted revisions and extensions are those made by Goddard, whose revision was first to appear in America, The Point Scale referred to above, by Yerkes, Bridges, and Hardwick, and the translation and revision by Kuhlmann.

It is beyond the scope of an elementary work of this kind to give a detailed discussion of the various revisions and criticisms of the Binet Scale. It is the purpose rather to give some of the more salient features of the attempts at revisions and criticisms and to refer the reader to a more exhaustive treatment of these things elsewhere.

The Stanford Revision of the Binet Scale.—To render the Binet Scale applicable to American conditions, some reorganization was necessary, first, because conditions in America were different from those in France, and second,

because it was felt that while the construction of the Binet Scale was fundamentally right yet there were a number of details that needed revision. According to Terman's view, many of the tests were misplaced. There was a dearth of tests at the higher mental levels; the procedure in giving the tests was not standardized; and many minor changes were necessary.

The revision was a long process involving several years of tedious work in examining and reexamining approximately 2,300 subjects, 1,700 of which were normal children, 200 defectives and superior children, and about 400 adults.

After Terman had compared the data from tests made with the Binet Scale in various parts of the world he decided to provide 40 additional tests to the Binet Scale to be used in the tryout for revision. This would make it possible to eliminate some of the least satisfactory ones and at the same time allow six tests for each age group instead of five. Care was taken to secure children whose ages did not vary more than two months from a birthday; that is, the age of eight-year-old group, for instance, ranged from seven years ten months to eight years two months. All the children within two months of a birthday were tested in order to avoid accidental selection. Tests of foreign-born children were eliminated in the final treatment of the data because they clearly did not represent a normal group. The children's responses to the tests were recorded almost verbatim. The revision of the scale below the 14-year level was based almost entirely on 1,000 unselected children. The object was to arrange the tests in such a way that the median mental age of the unselected children of each age-group would coincide with the median chronological age. For example, a correct scale must cause the average child five years old chronologically to test five years old mentally, and the

average six-year-old child to test six years mentally and so on.

This relation was expressed in what is known as the *intelligence quotient* discussed in the previous chapter. The scale above the 14-year age-level was based on the results of testing adults, since children in the grades older than 14 would be classified as retarded. The validity of particular parts of the scale was determined by dividing the children of each age-level into three groups according to intelligence quotients: (1) those testing below 90; (2) those testing between 90-109; and (3) those with an intelligence quotient of more than 110. The percentages of passes at or near that age-level was then ascertained separately for the three groups. If a test failed to show a decidedly higher proportion of passes in the superior group than in the inferior group, it was discarded as unsatisfactory.

The scale when finally completed consisted of 90 tests, 36 more than were included in the Binet 1911 Scale. There are six tests for each age-level from 3 to 10, eight at 12, six at 14, six at "average adult" age, six at "superior adult" age, and 16 alternative tests. The alternative tests are to be given only when some of the regular tests have been rendered unfit because of coaching or for other reasons.

A comparison of this scale with the 1911 edition of the Binet Scale above the adult group shows that two tests are eliminated and 29 are relocated, of which 25 are moved downward and four upward. Eighteen of the 29 relocated tests were moved down one year, four were moved down two years, two were moved down three years, one was moved down six years, three were moved up one year and six moved up two years.

To Find the Mental Age of a Child By the Stanford Revision.—Since there are six tests for each age-group

from III to X, each test passed counts two months towards mental age. For instance, if a child were to pass all the tests in the six-year group and two in the seventh, he would be considered six years and four months 'old' mentally. Since there is no 11-year group in the Stanford Revision and there are eight tests in the 12-year group, these tests must cover a period of 24 months and each test passed should add three months to a child's mental age.

The Picture Completion Tests.—A notable attempt to extend the Binet Scale is found in the picture completion tests devised by Healy.² His object was to test intelligence and at the same time eliminate the language factor.

Since modifications of the Ebbinghaus Completion Method are now quite extensively used for language and seem to correlate very highly with well-known tests of general intelligence, it seemed reasonable to Healy to suppose that practically the same sort of ability would be required to complete a picture as to complete a sentence. He therefore devised the Picture Completion Test.

In giving these tests, the essential thing for the pupil is to see that something is lacking in the general situation in the picture. The child is asked to supply the missing parts to complete the scheme. In the Picture Completion Test, the choice of a missing part is limited to blocks supplied to the subject, whereas in the language-completion tests in general use, the subject has the entire range of his vocabulary from which to supply the missing word.

The material for the Picture Completion Test consists of a brightly colored picture 10 by 14 inches in dimensions. It represents a barnyard scene in which ten simple activities are going on. There is no obvious connection

² Rudolf Pintner and Margaret M. Anderson, *The Picture Completion Test* (Warwick and York, 1917), ch. ii.

between the activities, but each is of such a nature as to appeal to the childish imagination. An object necessary for the completion of any one of the activities is omitted and it becomes the task of the examinee to find the most appropriate object. For example, two boys are playing with a football. One of the boys has just kicked it into the air and the other boy has his hands in a position for catching it, but the significant thing, the football, has been omitted and a blank space instead of the football appears between the two boys.

In addition to the ten appropriate blocks for completing the pictures there are 40 others from which the subject may choose, ten of which are blank while the other 30 bear pictures of objects. The picture board contains 10 apertures each of which is one inch square, and the blocks are so cut that any block will fit any aperture. No indication of the correct solution may be gained from the background of the picture; but the subject must grasp the meaning in order to meet the requirements.

A test of this kind offers some measure of a child's apperceptive ability and shows how well he is able to use his past experience in the solution of new and novel situations. This factor corresponds in the main to the definitions of intelligence given by Stern, Burt, Binet, and others. It is claimed that this test differs from the ordinary picture-puzzle tests inasmuch as it demands a choice on the part of the child. Healy claims the tests correlate well with the apparent mentality of both the delinquents and the normal individuals. The most of the work with this test has been done in corrective and protective institutions.

The Form Board for Measuring Intelligence.—The *form board*, which in some respects is very much like Healy's Picture Completion Test, was originally devised by Seguin. A false conception of psychology, the old faculty

psychology, was responsible for its creation, but it has since proved to be of value in mental diagnosis. Owing to the fact that Seguin believed defective children different from normal children in kind of mentality, he thought they would demand a different kind of test to measure their intelligence. He believed that general improvement could be brought about by training in specific tasks; that the mind could be divided into separate entities such as attention, will, memory, imagination, and the like, and that training in each of these gave a general training of the whole mind.

The form board, which consists of a board usually about 14 by 20 inches in dimensions, has a number of irregular-shaped apertures in it with a number of blocks that will fit the various apertures. Each block will fit but one aperture and the test consists in determining how quickly pupils will assemble the blocks into their appropriate apertures. Feeble-minded people have little sense of form and proportion, hence they must try many times before they can find the right blocks for the various apertures.

A Scale of Performance Tests.—Pintner and Paterson have designed and assembled a group of tests with the idea of making a scale which will supplement the intelligence scales now in use. Their scale is the result of an attempt to measure the mentality of deaf children and had its beginning in 1914. Not only is it desirable to measure the mentality of deaf children, but there are many other types that cannot be measured by the ordinary intelligence scales. The speech defective, the backward child, the foreign child, and many types of subnormal children do not respond to these tests in such a way as to reveal their mentality. A battery of performance tests, therefore, was thought by these men to be the *sine qua non* to measure their general intelligence. The term "performance tests" is used here in a restricted

sense and indicates a group of tests which involves a great deal of manipulation with the hands and a minimum amount of language responses.

The scale as reported by the authors consists of a group of fifteen tests as follows:³

1. The Mare and Foal Picture Board, a modification of the original as designed by Healy.
2. The Seguin Form Board, Twitmeyer's adaptation of the Goddard or the Goddard Board itself.
3. The Five-Figure Board, devised by Paterson.
4. The Two-Figure Board, devised by Pintner.
5. The Casuist Form Board, a copy of the original board devised by Knox.
6. The Triangle Test, devised by Gwyn.
7. The Diagonal Test, devised by Kempf.
8. Healy Construction Puzzle A, devised by Healy.
9. The Manikin Test, devised by Pintner.
10. The Feature Profile Test, devised by Knox and Kempf.
11. The Ship Test, devised by Glueck.
12. The Picture Completion Test, devised by Healy.
13. The Substitution Tests, devised by Woodworth.
14. The Adaptation Board, devised by Goddard.
15. The Cube Test, devised by Knox and modified by Pintner.

A brief description of the first test will give an idea of the general nature of the series.

The Mare and Foal Picture Board is a board measuring 29 by 24½ centimeters and one centimeter thick upon which a colored picture is pasted. The picture represents a mare and foal in a field with two sheep lying on the ground and three chickens in the foreground. In the background two houses are seen in the distance. Eleven pieces of irregular shape have been cut from the picture. Each piece represents certain parts of the animals or of

³ Rudolf Pintner and Donald G. Paterson, *A Scale of Performance Tests* (D. Appleton Co., New York, 1917), pp. 23-24.

the scene. In giving the test the board is placed in front of the child with the pieces scattered over the top. The instructions are to put the pieces in place as quickly as possible without making any mistakes. The examiner watches the performance of each child and records the number of errors made. An error is an attempt to put a piece in the wrong place. The time is five minutes.

Although the other 14 tests differ from this one, yet the general plan is the same. In selecting the tests for this scale, the aim was to obtain as many kinds as possible so that the various factors entering into the complex known as intelligence might be brought into play. Care was taken not to select the performance of a specific activity that was likely to have been *learned* by the child. The tests must be of such a nature that no verbal instructions are necessary.

A Point Scale for Measuring Mental Ability.—There have been many differences of opinion as to whether the "all-or-none" method of scoring used in the Binet Scale was yielding the most satisfactory results. Many psychologists have been convinced that a more nearly perfect picture of a child's mentality might be drawn if the scale were a little finer; that is, if the units of accomplishment were made smaller so that a child would get credit for any part of a test properly completed instead of having to complete the whole test in order to score.

With this idea in mind, Dr. Robert M. Yerkes, assisted by James W. Bridges and Professor Rose S. Hardwick, undertook the task of revising the method of scoring used in the Binet Scale. Their intentions were at first simply to devise a better method rather than to attempt to modify the Binet Scale. A number of preliminary tests were given to approximately 1,000 school children and inmates in a psychopathic hospital to determine the value of the

single-series and the partial-credit ideas before attempting to develop a more highly satisfactory form of point scale. The main defect they sought to correct in the Binet Scale may be illustrated from the following example taken from the Binet Scale.

Suppose three pupils, *A*, *B*, and *C*, were asked to count backwards from 20 to 0. *A* makes the count without error, *B* makes two errors, and *C* fails entirely. In recording the results by the Binet method, *A* is given a perfect score, and *B* and *C* are recorded simply as failures and are put in the same class. Now it is evident that *B*'s score is much nearer like *A*'s score than *C*'s. The point scale would give *A* a perfect score, *B* a certain number of credits, and *C* a score of zero, which would be a more equitable rating of the three abilities.

It is claimed also that the point system gives due credit to the more difficult reactions which are many times not properly weighted by the other method. By making the units of accomplishments smaller, it is possible to note gains made in case the test is repeated. In the case cited above, if the test were again given to *C* and he was able to count from 20 to 0 with only four mistakes, his gain would be noted, but by the "all-or-none" method no progress would be recorded until he was able to make the count without error. It is claimed by the authors that the point method of scoring tends to minimize the influence of the personal equation of the examiner and, in doubtful cases where the examiner is not quite sure whether the examinee made a perfect score or not, the pupil may be more accurately graded.

It is argued that the ideal examination question is one upon which the abler students will make a high score and on which the poorly prepared will be able to give some answer even though the answer is not of such a nature as to merit a perfect score. The object is not simply to

know that certain individuals have a passing knowledge of the topic, while certain others have not, but to know how much the better candidates surpass the minimum requirement and to what degree the less able students fall short of it. A point scale opens the way for the classification of individuals in more nearly homogeneous groups than they may be classified under the "all-or-none" method. The dominating idea of the whole scheme is to *take cognizance of small gains*; to make small units of accomplishment, rather than large ones, the units of measurement. The makers of the Point Scale also introduced the *intelligence coefficient* which is obtained by dividing the number of points obtained in an examination by the number of points obtained on the average. For example, if 40 points were the average for eight-year-old children and a child were to make 36 points, his intelligence coefficient would be $36 \div 40 = 0.9$. A similar measure was suggested earlier but had not been put into practice prior to the time of the Point Scale.

The testing material for the Yerkes, Bridges, and Hardwick Point Scale was drawn very largely from the material used by Binet. In fact, as has been said, the original intentions of the makers of the Point Scale were to develop a better method but not to attempt to modify the Binet Scale. But as the work progressed the makers were convinced that much of the material in the Binet Scale was inadequate, and changes were made where it was deemed necessary. The tests were selected with the intention of covering the various common forms of the principal mental functions.

Tables I and II on the pages following show the distribution of the tests used among these principal mental functions.⁴

⁴ *A Point Scale for Measuring Mental Ability*, pp. 7-9.

TABLE I

Tests

1. Auditory memory for sentences, attention.
2. Perception (visual—of things, relations, meanings), apperception, association, imagination.
3. Auditory memory for words (digits), attention.
4. Discrimination—(a) visual, (b) and (c) kinesthetic.
5. Motor coördination, visual perception.
6. Ideation (association and analysis).
7. Aesthetic judgment involving perception, association, and analysis.
8. Perception, apperception, visual memory, imagination.
9. Association (free), vocabulary, attention.
10. Analysis and comparison of remembered objects, attention.
11. Memory, imagination, attention.
12. Practical judgment involving memory and imagination.
13. Kinesthetic discrimination, ideation (notion of series), attention.
14. Imagination and command of language forms.
15. Logical judgment based on imagination, analysis, and reasoning.
16. Suggestibility, visual perception, comparison.
- 16a. Logical judgment based on analysis and reasoning.⁵
17. Ideation involving vocabulary, memory, analysis.
18. Logical judgment based on analysis and reasoning, attention, memory.
19. Visual memory, perception, attention, motor coördination.
20. Ideation involving analysis, imagination, command of language forms.

Some of the principles for the selection of testing material were as follows: Other things being equal, preference was given to tests applicable through a considerable range of years, such as memory span and free association; and the different reactions to a given test which are characteristic of successive stages of mental growth were dis-

⁵ This is an extra test, introduced as a possible substitute for test 16 at a time when that seemed likely to prove unsatisfactory.

TABLE II

Mental Processes	Tests	Credits
Motor coördination.....	5	4
Perception (visual).....	2, 8	13
Discrimination (visual).....	4a	1
Discrimination (kinæsthetic).....	4 b and c, 14	4
Association.....	9	4
Suggestibility.....	16	3
Memory.....	11	4
Memory (auditory).....	1, 3	11
Memory (visual).....	19	4
Imagination.....	13	4
Judgment (æsthetic).....	7	3
Judgment (practical).....	12	8
Judgment (logical).....	15, (16a), 18	11
Analysis and comparison.....	10	6
Ideation.....	6, 17, 20	20

criminated in the scoring wherever easily recognizable.* For example, four gradations are recognized in the free association test, two in the definition of concrete terms, four in counting backward, and so on.

Aside from the points mentioned above the makers of the Point Scale state further reasons for the superiority of their scale over the Binet Scale. It is committed to no hypothesis as to the correlation existing between chronological and mental age, or between the different mental functions at different stages of development. It is capable of giving results of ever-increasing reliability and precision as data accumulate and norms are established; it works with a smaller amount of testing material, which makes it possible to select better material. The Point Scale consists of 20 tests which include about 65 questions, whereas the Binet pre-adolescent scale consisted of 52 tests with approximately 100 questions.

Though the scale described above is for children, its

* *A Point Scale for Measuring Mental Ability*, p. 9.

makers present a list of principles for a universally applicable scale for mental ability that will include adults as well. Among these principles are the following:

4. Distribution of the several measurements in the series equally among the chief mental processes, as, for example, according to the four following categories:

(a) Receptivity, including such functions as sensibility, perceptivity, discrimination, and association.

(b) Imagination, including memory, in its various aspects, and constructive imagination.

(c) Affectivity, including simple feeling, emotion, sentiment, volition, and suggestibility.

(d) Thought, including ideation, judgment, and reasoning.

5. Selection of twenty parts of the scale so that there shall be five for each of the groups of mental functions, classified under the headings receptivity, imagination, affectivity, and thought.

6. A minimum of 200 points, one-fourth of which shall belong to each of the above-mentioned groups of processes.

The individual's score is to be in terms of the four mental processes instead of throwing the scores all together and a norm to be determined for all mental process.

The two underlying principles of the Binet Scale are: first, the arrangement of the tests in groups corresponding to the years of chronological age and the consequent expression of the results as "mental age," and second, the related "all-or-none" method of scoring. Many objections are raised against each of these principles. In reference to the first principle, which assumes that all normal individuals develop mentally by similar stages, that the correlation between the different functions is the same for all individuals at a given stage, and that physical and mental development correspond, the critics of the scale claim that these statements are not warranted by the facts. On the contrary, it is pointed out that such studies as those made by Decroly and Degand in Belgium show that on the average the Belgian children are a year and a half in advance

of the group selected by Binet at Paris. Binet's attempt to account for this discrepancy on the ground that the Belgian children belong to a more privileged class is not accepted as scientific.

Binet's tests are criticized also on the ground that a range of six or seven years is included within what constitutes the "normal age." Binet said that the children in one quarter of Paris were found to be advanced by four or even five years, and adds that "one must, therefore no longer consider the retardation or advance of three years as an anomaly." This statement is criticized on the ground that this is a very large proportional variation for a scale that covers at most but twelve years.

In scoring the examination, the examinee is credited with that age at which he passes all the tests, plus one year for every five tests passed from more advanced groups. The more difficult tests designed for advanced ages do not count any more than those lower down when the final score is being reckoned. This seems to be an inequitable distribution of credit because more credit should be given for tests passed in the more advanced age-groups than for tests passed in the lower age-groups.

Proposed Reorganization of the Binet Scale by Meumann.—Meumann⁷ has indicated the lines along which the Binet Scale should be reorganized and extended. His proposals for reorganization fall under three heads as follows:

I. *Endowment or Intelligence Tests Proper.*—Under this heading he includes tests on:

1. Concentration and fixation of attention.
2. The immediate retention of verbal and non-verbal material.
3. Imagination and thinking by means of descriptions of pictures.

⁷ *Vorlesungen*, Vol II, pp. 286-288.

4. The combination test improved by Ebbinghaus.

5. Employing and concentrating on visual images in special problems, as in the folded paper test, or such a test as: What kind of a solid figure is produced when a right triangle is rotated around one of its sides making the right angle?

6. Thinking by means of controlled association test.

7. The concentration and synthesis through comparisons and differentiations of different difficult objects both in the presence of the objects and in recollection.

II. *Tests of Development in the Narrower Sense.*—Here would be given:

1. Vocabulary tests.

2. Tests measuring the range of attention and the memory span.

3. Tests determining the examinee's ability to reproduce words, as in the opposite test.

4. Tests in temporal orientation, and so on.

III. *Tests of Environment.*—These tests may be divided into three divisions:

1. Testing school knowledge, as knowledge of number, coins, days of week, dates, names of months, writing from a copy, and writing from dictation.

2. Knowledge acquired at home; meaning of words designating household articles, above and below, before and after, left and right, correctness in speech, testing errors in speech, number of the fingers, and so on.

3. Spontaneous observation, by making inventories of mental content.³

Wallin's Criticisms.—Wallin criticizes the 1911 edition of the Binet Scale on the ground that the number of tests in each group should have been increased rather than diminished and a greater number of functions covered. In regard to the principle of an age-scale, Binet's critics claim he surrendered its validity on many occasions, as,

³ Cf. Robert R. Rusk, *Experimental Education* (Longmans, Green & Co. 1919), pp. 185-187.

for instance, when he accounts for Belgian children testing higher than those of Paris because of their having a more favorable environment; also when he takes exception to Katherine Johnson's results because she treated as a single group children from different levels of privilege; and again when he tells of the wide range among normal children.

Other Tests Devised to Measure Intelligence.—Many attempts have been made by still others to devise tests to measure intelligence. Some have attempted to improve the Binet tests while others have devised entirely new ones. The following have attracted considerable attention.

Burt⁹ has attempted to improve the method of recording the mental development of children. Instead of using the mental age as a measure for backwardness and mental deficiency, he has used the standard deviation (for a definition of standard deviation see Chapter XI) of normal pupils, that is, the pupils who are in the usual class in school for their chronological age. Taking the standard deviation for the various grades he found it to be about one-tenth of the chronological age. That is, the standard deviation of a normal child 10 years old is one-tenth of his chronological age, or one year. A child five years old would have a standard deviation of a half year and one 15 years old would have a standard deviation of one and one-half years.¹⁰ He found that by taking the standard deviation as a measure a child who was retarded by more than three-tenths of his age needed a special school for his training. He says:¹¹

... for practical purposes, "backward" may be taken to denote children who, though not "defective," are yet unable, in

⁹ C. Burt, *The Distribution and Relations of Educational Abilities* (King and Son, London).

¹⁰ *Ibid.*, p. 31.

¹¹ *Ibid.*, p. 82.

will, apprehension, visual memory, memory for the elements of speech, numerical memory, comprehension, combination, mechanical sense, imagination, and observation. All of these have several subdivisions. There are ten tests for each process or partial process and the score for each process is indicated by a mark on a table. The points are then joined up and the whole figure results in what is known as a "mental profile."¹² The score for each process is indicated by the mark on the table.

De Sanctis devised a series of six tests primarily to measure not the level of intelligence, as do the Binet tests, but the degree of mental defection.

A general idea of the de Sanctis tests may be gained from the description given below:

1. Five balls each of a different color are placed before the subject and the examiner says: *Give me a ball.* The time and manner of the response is noted.

2. The same five balls are arranged in a row and the examiner says: *Which is the ball you just gave me?* The time of the response is again noted.

3. The subject is shown a wooden cube such as is used in a kindergarten and the examiner says: *Do you see this block of wood?* After the subject has noted it, the examiner continues: *Pick out all the blocks like this on the table.* On the table have been placed three cones, five cubes and two parallelopipeds. The time for selecting and arranging the cubes is noted.

4. The examiner shows the subject a cube and says: *Do you see this block?* After noting the block the examiner continues: *Point out a figure on the form chart that looks like it.* The form chart consists of ten rows of squares, triangles and rectangles with fourteen figures in each row, or 140 figures in all. If the subject points to a square, the examiner's next command is: *Take this pencil (or pointer) and point out all the squares on the chart as fast as possible, without missing any, taking the figures line by line.* The time, mistakes, and omissions are noted.

¹² "Mental Profiles: A Quantitative Method of Expressing Psychological Processes in Normal and Pathological Cases," *Journal of Experimental Pedagogy*, Vol. 1, pp. 211-214.

5. Additional blocks are placed on the table in such a manner that the distance between the cubes is not more than two centimeters. Each cube should be just one-half centimeter longer on each side than the next smaller one. If it is desired to make the test more difficult, the number of cubes may be increased or the difference in size decreased. The examiner then says: *Here are some more blocks like those you have pointed out on the chart. Look at them carefully and tell me: (1) How many there are. (2) Which is the largest. (3) Which is the farthest away from you.* The time, errors and omissions are noted.

6. The examiner asks: *Do large objects weigh more or less than small objects? Why does a small object sometimes weigh more than a large one?* The second question is asked if the first one has been answered correctly. The subject is then asked: *Do distant objects appear larger or smaller than near objects? Do they only seem smaller or are they really smaller?* (The last question will show whether the subject is aware of optical illusions.)

De Sanetis points out that if the subject does not pass the second test the mental deficiency may be considered of a high grade. If the subject cannot go beyond the fourth test, or if he makes many mistakes, or is not at all certain in the fifth, the mental deficiency may be considered slight. If the sixth test is completed without a mistake the subject may be said to present no mental deficiency.

II. GROUP INTELLIGENCE TESTS

(One of the greatest drawbacks to the Binet Scale from an administrative standpoint is the fact that the tests must be given to but one individual at a time. This makes their administration a great burden if the number of persons to be given the tests is very great. In order to overcome this difficulty, group tests for mental ability have been introduced. The gain from an administrative standpoint, however, has not been an "unmitigated blessing," since much had to be sacrificed in order that the test might be given to groups of children.

On the other hand, the group intelligence tests were used in the American army with marked success. The experience with the drafted men in the army proved that not only individual tests of intelligence, but group mental tests could be used to good advantage. The tests were thus introduced into the army, and psychological methods were used in the selection of its personnel.

The American army was the only army that made use of intelligence tests in the selection of officers. As a result of the application of the group intelligence tests to hundreds of thousands of men, and individual tests to a very large number of the same men, correlations were made between the two test series, and between the single tests and various other criteria for judging intelligence. By these methods the group tests were found to be very reliable.

For the army Beta group mental tests, no educational training was necessary, not even the ability to read and write. The tests were given to illiterate foreigners who did not know the English language and also to deaf mutes. Some of the illiterate foreign recruits made scores higher than the average of the commissioned officers.

For the army Alpha group intelligence tests, ability to read and write and to understand English is essential; therefore, education through the third and fourth grades is probably necessary, but after this point, educational training has little effect upon the results. As a matter of fact, some with almost no educational training made among the highest scores.

Principles Involved in the Selection of Group Tests.—The first requirement in the selection of any test is that its beginning be easy enough and the directions clear enough that all the children may be able to do a part of the test.

One of the purposes for devising and standardizing

group tests is to provide a scale for measuring the intelligence of large groups of children with sufficient accuracy to sort out all children of questionable mentality. Another purpose is to obtain a group scale that will discriminate between dull, average, and supernormal children. When group tests are given one can be reasonably certain that children who make exceptionally low scores will, when tested individually by the Binet revisions, be found to be mentally deficient. This method saves much time.

Requirements of Group Tests Are Many.—They must not only possess the characteristics necessary for the individual tests but satisfy other demands. Simplicity is an important criterion for the selection of group tests; simplicity of material, of directions, of response, and scoring. Since a large group of children are to be tested at one time, it is necessary that the tests be selected in which the material used can be easily carried and quickly distributed.

Simplicity of directions is one of the very essential characteristics of group tests for the lower grades, especially those below the third. The following experience recorded by Miss Frances Lowell illustrates the point in question. She says:

Frequently one feels confident that the directions for a certain test are perfectly clear and that they could not possibly be misunderstood, and yet when the test is given to a group of children he finds that the meaning is entirely lost. Good English must frequently be sacrificed, for the child in the primary grades is surprisingly limited in vocabulary. Originally in giving the directions for the one of the six-year tests the writer said: "Make a cross in the *largest* square." The result was puzzling, for the children crossed the smallest square as often as they did the largest. The test apparently was a failure for that age-group. However, before discarding it the writer decided to experiment a little to see if the difficulty could be discovered. Instead of having

the children cross the square they were asked to point to the largest square, whereupon one little girl tearfully informed the writer that she didn't know what that meant. That solved the problem; from then on children were instructed to make a cross in the *biggest* square.

The Terman Group Tests of Mental Ability.—Dr. Lewis M. Terman devised a series of group tests for grades 7 to 12 inclusive. Like the Binet tests they consist of a series of questions and problems selected from a large mass of possible test material. Each separate item was correlated with a dependable measure of mental ability. The preliminary trial series from which the tests were finally made was composed of a two-hour test of 13 parts with 886 different items. As a result of the preliminary tryout, three of the 13 parts were eliminated and the number of items in the remaining ones reduced to 370. If an item failed to differentiate pupils of known brightness from pupils of known dullness, it was eliminated. Much attention was given to simplicity and convenience in making up the tests. The tests are issued in two forms, A and B. Each contains 185 questions or problems. In any test of intelligence there is always a margin of error. The author of this test suggests that both forms A and B be used and that the average result obtained from both forms be used as a possible basis for guidance of the individual pupil. Each form consists of 10 groups of tests. Each group has a number of questions or problems to be solved.

TEST 1. Information.—It consists of 20 parts. The pupil is told to draw a line under the one word that makes the sentence true. *EXAMPLE: Coffee is a kind of bark, berry, leaf, root.* Time, two minutes.

TEST 2. Best Answer.—It consists of 11 parts. The examinee is asked to put a cross before the best answer to the question or statement made. *EXAMPLE: Spokes of a wheel are often made*

of hickory because: (1) hickory is tough, (2) it cuts easily, (3) it takes paint nicely. Time, two minutes.

TEST 3. *Word Meaning*.—It consists of 30 pairs of words arranged as follows: *Expel — retain — same — opposite*. The instructions are: "If the words mean the same, the examinee is to underscore the word *same*, if they mean the opposite, underscore the word *opposite*. Time, two minutes.

TEST 4. *Logical Selection*.—The instructions are to draw a line under two words that the thing *always* has; underline *two* and *only two* in each line. EXAMPLE: *A horse always has harness, hoofs, shoes, stable, tail.*

TEST 5. *Arithmetic*.—This test consists of 12 problems of which the one given below is a sample: "How many hours will it take a person to go 66 miles at the rate of 6 miles an hour?"

TEST 6. *Sentence Meaning*.—It consists of 24 parts. Instructions: "Draw a line under the right answer." EXAMPLE: *Does a conscientious person ever make a mistake? Yes, No.*

TEST 7. *Analogies*.—Consists of 20 sentences in which the analogies are to be picked out. EXAMPLE: *Ear is to hear as eye is to table, see, hand, play.* The examinee is required to underscore the proper word. Time, two minutes.

TEST 8. *Mixed Sentences*.—Consists of 18 sentences with words not arranged in proper order. The instructions are to arrange the words in correct order and then tell whether the sentence is true or false by underscoring one of these words. EXAMPLE: (1) *True bought cannot friendship be — true, false.* Time, three minutes.

TEST 9. *Classification*.—Consists of 18 groups of words, one word in each group of which does not belong to that class. The instructions are to cross out the word that does not belong there. EXAMPLE: *Automobile, bicycle, buggy, telegraph, tram.* Time, three minutes.

TEST 10. *Number Series*.—Consists of 12 rows of numbers arranged in such a way that the examinee can predict what the next two missing numbers should be by the arrangement of the numbers that are given. EXAMPLE: *3 — 8 — 13 — 18 — 23 — 28 — ? — ?*

Examination Form B is constructed on the same plan as Form A and is of approximately the same difficulty.

The scoring is done by ten scoring keys which, for convenience, are printed on a single sheet.

Mental ability should be the fundamental basis for all grading promotion and classification of pupils. Terman emphasizes the fact that the purpose of these tests is *"to make a difference in the educational treatment of pupils, not to gratify a merely idle curiosity regarding their intellectual status."*

Accordingly, one of the chief functions of these tests is in the classification of pupils into different groups for special types of instruction, as in the Oakland Plan, which divides the pupils into three groups—the bright, the average, and the slow, for separate instruction according to their needs.

The use of the tests will clear up many misunderstandings regarding individual children. They help to give reasons for the seeming misfits in school. They give educational guidance that helps to simplify the problems of vocational guidance. It is not claimed, of course, that the tests are infallible, but it is claimed that they make a very important point of departure for further study of the pupil.

The National Intelligence Tests.—In March, 1919, the General Education Board granted the National Research Council the sum of \$25,000 to be used in devising methods for measuring the intelligence of school children. The Research Council caused to be organized a committee to do this work. The committee was composed of Messrs. R. M. Yerkes, chairman, M. E. Haggerty, L. M. Terman, E. L. Thorndike, and G. M. Whipple. *Martin*

The committee decided to prepare a series of tests for children from grades 3 to 8 inclusive. From a great mass of data, material was selected for 21 tests for a preliminary trial. Children in a number of eastern cities, including Washington, D. C., Cleveland, New York, Richmond,

and Alexandria, were given the preliminary tests. Dr. Truman Kelley statistically analyzed the data obtained from this preliminary trial. Guided by Dr. Kelley's report and other information relative to the characteristic behavior of the several tests, the committee selected 10 tests from the list of 21 for further use.

These ten tests were arranged in two groups, or series, of five tests each. The tests selected are an adaptation for school purposes of the group intelligence tests used in the examination of recruits in the United States Army.

Each scale is a complete unit for testing; but it is recommended that both scales be used in order that one may serve as a check on the other. A short preliminary exercise is given before each regular test to acquaint the pupil with the general nature of the work to be done. The tests in Scale A consist of:

TEST 1. *Arithmetical Reasoning*.—This test consists of 16 problems growing progressively more difficult from the first to the sixteenth. The time allotted to this test is five minutes. The instructions given the pupils are to solve as many problems as they can in the time allotted.

TEST 2. *Sentence Completion*.—This test consists of 20 sentences, parts of which have been left out. The instructions are to fill the blanks with words to make the sentences sound sensible and right. Time, four minutes.

TEST 3. *Logical Reasoning*.—Consists of 24 parts. Part, or row, 1 is given here to indicate the general nature of the test:

Elephant (circus, ears, hay, keeper, trunk)

The instructions are: "In each row draw a line under each of the two words that tells what the thing always has." Time, three minutes.

TEST 4. *Same-Opposite*.—Consists of 40 pairs of words separated by a blank line, as *new—old*, *still—noisy*, *fall—drop*, etc. The instructions are to write "S" between them if they mean the same and "D" if they mean different things.

TEST 5. *Symbol-Digit*.—Consists of nine irregular-shaped characters arranged in six rows with 20 characters in a row. The nine digits, 1, 2, 3, etc., are placed in a line under the nine characters, each digit corresponding to a character. This is called the key. The instructions are to make under each drawing or character the number you find under the drawing in the key. Time, three minutes.

Scale B consists of five tests as follows:

TEST 1. *Computation*.—This test is made up of 22 problems in arithmetic. The problems grow progressively more difficult as the pupil works down the page. Time, four minutes.

TEST 2. *Information*.—It consists of a series of 40 sentences, each sentence containing four words, only one of which makes the sentence true. Sentence Number 1 reads as follows: *The day before Sunday is Friday, Monday, Saturday, Tuesday*. The instructions are to draw a line under the one word that makes the sentence true. Time, four minutes.

TEST 3. *Vocabulary*.—Forty questions to be answered by "yes" or "no" are given. Time, three minutes.

TEST 4. *Analogies*.—This test consists of 32 parts, each part consisting of seven words. The first two words bear a definite relation to one another. The third word bears an analogous relation to one of the four succeeding words in the line. Part 1 of this test is given here to illustrate the point: *Finger—hand—toe, box, foot, doll, coat*. The instructions are: Read the first three words in each line. Then read the last four words and draw a line under the right one.

TEST 5. *Comparison*.—The test consists of 50 figures, names, and drawings, arranged in pairs, with a dotted line between them. The instructions are: "If the two things in a pair are the same write 'S', if they are different write 'D.'"

The Haggerty Intelligence Examinations.—Dr. M. E. Haggerty devised a series of intelligence tests known as Intelligence Examinations, Delta 1 and Delta 2. Delta 1 consists of six groups of tests designed for children of grades one to three. Delta 2 is composed of the same number of tests and is designed for children in grades

three to nine. Each of these tests is preceded by a fore-exercise which gives the pupil an idea of what he is to do in the regular test.

The material in Delta 1 is of such a nature that a child unable to read may do several of the exercises. The first regular test is one of *oral directions*. It consists of a page of pictures and depends on the oral directions of the examiner to do the test. For instance, the first picture is a mouse and the oral directions are to draw a ring around the mouse.

TEST 2. *Copying Designs*.

TEST 3. *Picture Completion*.—A part of the picture is left out and the child is asked to supply the missing parts.

TEST 4. *Picture Comparison*.—Here each line contains two pictures separated by a blank line. If the pictures are the same the pupil writes "S" on the blank line; if they are different he writes "D."

TEST 5. *Symbol-Digit*.—The nature of the symbol-digit test was explained in the discussion of the National Intelligence Tests.

TEST 6. *Word Comparison*.—This is the first of the series that requires a reading ability on the part of the examinee. This test is like the word comparison test described above.

Delta 2 consists of six tests as follows:

TEST 1. *Sentence Reading*.—It consists of a series of 40 questions growing progressively more difficult that may be answered by "yes" or "no." Time, five minutes.

TEST 2. *Arithmetical Problems*.—A list of 20 problems is given and the pupil is asked to solve as many as he can in five minutes.

TEST 3. *Picture Completion*.—Time, four minutes.

TEST 4. *Synonym-Antonym*.—Consisting of 40 pairs of words. This test is the same as the same-opposite test described above.

TEST 5. *Practical Judgment*.—It consists of a series of 16 sentences and questions, each of which has three answers, or reasons for being, one of which is right. The child is asked to put a cross before the best answer to the statement or question.

TEST 6. *Information*.—Consists of 40 sentences. It is similar to the one described under the National Intelligence Tests.

The Otis Group Intelligence Scale.—The Otis Group Intelligence Scale is designed to test general mental ability. The scale is issued in two series, a Primary Examination and an Advanced Examination. The Primary Examination is designed especially for the kindergarten, and for grades 1 to 4, and consists of eight tests which do not involve the ability to read. The Advanced Examination, consisting of ten tests is designed for grades 5 to 12; in fact, for all literary persons, including university students.

In each series, each test is independent of every other test, but the tests taken together form a scale, which is printed in the form of an examination booklet. The advanced examination consists of ten tests, each test consisting of a series of questions and problems. The tests are as follows: (1) *following directions* (2) *opposites*, (3) *disarranged sentences*, (4) *proverbs*, (5) *arithmetic*, (6) *geometric figures*, (7) *analogies*, (8) *similarities*, (9) *narrative completion*, and (10) *memory*. This advanced examination is suitable for testing all students from the fifth grade upward through the high school and the university.

The Otis tests are very popular and are praised very highly by others who have designed group tests of intelligence. Terman writes as follows about the Otis Advanced Examination:¹³

It is applicable to any individual, whether child or adult, who has had the equivalent of three or four years of schooling. With subjects of this amount of schooling the Otis Scale probably comes as near testing raw "brain" power as any system of tests yet devised. Indeed, it was the first scientifically grounded and satisfactory scale for testing subjects in groups. . . . No one else has done so much as Dr. Otis to free intelligence tests from

¹³ *The Otis Group Intelligence Scale, Manual of Directions*, p. 3.

the influence of the personal equation of the examiner. Perhaps it is too much to hope that any mental tests can be made "fool proof" but it is not too much to say that the Otis Scale can be correctly given and correctly scored by any one who is intelligent enough to teach school. The plan of arranging the tests so that they may be scored by the use of stencils is a contribution of both practical and scientific importance.

The Dearborn Group Intelligence Tests.—Professor Dearborn has designed two series of tests, one for grades from 1 to 3, and the other for grades 4 to 9 inclusive. Series I consists of general examination 1, 2, and 3. Series II consists of general examinations 4 and 5. Though these tests differ from the other group tests described above, they are very similar to them in many respects and involve practically the same mental functions as the other tests.

Uses of Intelligence Tests.—The uses of intelligence tests are many and varied. They are beginning to play an important part in determining vocational fitness. No one will claim, of course, that they will tell us unerringly for which of a thousand or more occupations an individual is best fitted. Sometime in the near future, however, industries will begin to set the minimum intelligence quotient for employees in their business. We now know that of the people who belong to the ranks of the industrial inefficient there is a very high percentage who are of sub-normal intelligence. Of 150 "hoboes" that Mr. Knollin tested a few years ago, 15 per cent belonged to the *moron* grade of mental deficiency and almost as many were borderline cases.¹⁴

A bulletin published by the War Department on Army¹⁵ mental tests shows the intelligence level for the various

¹⁴ Cf. Lewis M. Terman, *The Measurement of Intelligence*, p. 54.

¹⁵ *Army Mental Tests, Methods, Typical Results and Practical Applications*, November 22, 1918, Washington, D. C.

THE MEASUREMENT OF INTELLIGENCE 133

occupations as determined by intelligence tests. The scores for some of the occupations are given below:

- 45 to 49.—Farmer, laborer, general miner, and teamster.
- 50 to 54.—Stationary gas engine man, house hostler, horse-shoer, tailor, general boilermaker, and barber.
- 55 to 59.—General carpenter, painter, heavy truck chauffeur, horse trainer, baker, cook, concrete or cement worker, mine drill runner, bricklayer, cobbler, caterer.
- 60 to 64.—General machinist, lathe hand, general blacksmith, brakeman, locomotive fireman, auto chauffeur, telegraph and telephone lineman, butcher, bridge carpenter, railroad conductor, railroad shop mechanic, locomotive engineer.
- 65 to 69.—Laundryman, plumber, auto repairman, general pipefitter, auto engine mechanic, auto assembler, general mechanic, tool and guage maker, stock checker, detective and policemen, toolroom expert, ship carpenter, gunsmith, marine engineman, hand-riveter, telephone operator.
- 70 to 74.—Truckmaster, farrier, and veterinarian.
- 75 to 79.—Receiving clerk, shipping clerk, stock keeper.
- 80 to 84.—General electrician, telegrapher, band musician, concrete constructor foreman.
- 85 to 89.—Photographer.
- 90 to 94.—Railroad clerk.
- 95 to 99.—General clerk, filing clerk.
- 100 to 104.—Bookkeeper.
- 105 to 109.—Mechanical engineer.
- 110 to 114.—Mechanical draughtsman.
- 115 to 119.—Stenographer, typist, accountant, civil engineer,
Y. M. C. A. secretary, medical officer.
- 125 and over.—Army chaplains, engineer officers.

With the educational levels of the various occupations thus determined, those attempting to give vocational guidance may have the sanction of science in the judgments they make. Suppose, for example, a pupil aspired to be a mechanical engineer. The intelligence level for

mechanical engineers is from 105 to 109. Suppose further that the intelligence level of this particular pupil is only 82. Then the vocational counselor may say to him that he is aspiring to do work on an educational level several units beyond the level for which his general intelligence fits him and that, if he persists in his attempts to be a mechanical engineer, he will find it extremely difficult to succeed because he does not have the potential power to succeed on that level.

There are many things that the vocational counselor must bear in mind when attempting to measure intelligence and determine the vocational fitness of individuals for the various businesses and professions. It may be, as Thorndike says, that we have three intelligences instead of one *general* intelligence, namely abstract intelligence, social intelligence, and mechanical intelligence. The intelligence tests we now employ are primarily designed to measure abstract intelligence, the ability to do abstract reasoning in arithmetic, logic, to deal successfully with abstract ideas, etc. An individual may make a high score with tests of this kind and yet score low in matters of social intelligence. The following example will make the matter clear: College professors, teaching mathematics, are usually considered to be good reasoners. Mathematics is supposed to bring into play the reasoning faculties of the mind. If one is a good reasoner in mathematics, will it follow that he will be able to reason successfully in a business situation which is pretty largely social? We know that there are many business men who would not score as high as the mathematicians on the intelligence tests now in use, but would be very much better in analyzing a business situation than the college mathematician. McCall points out that there is a much higher correlation within any one of these intelligences than between any two of them.

If these hypotheses are true, then the vocational expert must be slow to assign an individual to a particular educational level unless he has tested him with tests that reveal his level in that particular intelligence.

The Binet tests found early and widespread use in juvenile courts, in state surveys for feeble-mindedness, and in many other fields of research. Each of the various fields has a literature of its own. It has been estimated that approximately 4,000,000 pupils were tested in the public schools within thirty months after the appearance of the first group intelligence scale. Children of the intermediate and grammar grades are the ones most favored in this respect. Classifications are being made on a basis of the scores made on these tests. It is not claimed that all the classifications made are wise and will lead to better results. In the main, however, the results are gratifying and have been worth while.

The workers in the field are growing more cautious as the work advances and are warning those who are less familiar with the limitations of the tests to be very conservative in their claims for them.

Wherever intelligence tests have been given in the schools, they have shown that approximately 2 per cent of the school population will never develop beyond the 11 to 12-year-old level. The tests are giving us much valuable information along many lines. Healy and others have pointed out that courts are in the habit of administering punishment to juvenile offenders without knowing very much about the mentality of those whom they direct. A very high percentage of the social offenders are mentally deficient. Mental testing is beginning to shed much light on these problems and less injustice will undoubtedly be done in the future in passing judgment on our social offenders.

The use and abuse of the tests have aroused widespread

criticisms, both constructive and destructive. In any event, the criticisms have given us a clearer understanding of what intelligence is and how to measure it.

III. SUMMARY AND EVALUATION OF THE MEASUREMENT OF INTELLIGENCE

In the foregoing pages we have noted some of the major problems that confronted those attempting to measure intelligence. We have also pointed out some of the attempts to solve these problems and have noted some of the more important tests and scales now in use. We shall now attempt to summarize and evaluate the movement to measure intelligence and bring our description of it up to date and also note the fields where special research work will probably be done in the immediate future.

We noted in the first part of Chapter III that the scales, or groups of tests to measure intelligence, have arisen from the individual tests. The mental scales are merely a grouping together of these individual tests in order to give a more general picture of the mental make-up of the individual. The first tests were concerned with the specific "faculties," capacities, or abilities of the mind. They came into existence when the old "faculty psychology" was still in good repute. It was then thought that the proper way to measure intelligence was to choose any faculty or trait of the mind for investigation and that the data obtained from the measurement of the trait chosen would be indicative of the general intelligence of the individual. It was also thought that these faculties, or abilities, functioned somewhat independently of one another and that they were easily isolated for investigation.

It was soon found, however, that the old "faculty psychology" idea was untenable and that, instead of a trait or capacity functioning more or less independent of

other capacities, it was very closely associated with them in its functioning and that the complete isolation of a mental trait for measurement was impossible.

All are agreed that there is a native capacity that conditions all mental functioning. Just what this capacity is and how it manifests itself are unsettled questions. It must not be inferred, however, that, because psychologists cannot define and completely measure this native endowment, we cannot turn the measurements thus far made to good account. Data obtained from measurements now made are not only of theoretical and scientific value but practical value as well.

We are just entering one of the most fertile fields of research that science will ever explore. Everything of a constructive nature that man does is conditioned and determined by this mental capacity we are attempting to measure. All philosophy, science, and art wait upon its growth and development.

Methods Are Yet Crude.—Just as the pioneers entering a new country must of necessity use crude methods until explorations are made and the needs of the country determined, so the scientists entering a new field, with strange environment, and with tools and equipment designed for other fields, must work at a disadvantage until he has orientated himself and has his bearings in the new field. The psychologists are just beginning to get their bearings in this new field of intelligence measurements. They are beginning to see that many of the implements used for measurements in other fields are ill-adapted to this new enterprise. They are also learning that certain other tools are indispensable in the explorations and measurements of mental traits and capacities. Tests involving merely visual acuity, such as the cancellation of the A's on a printed page, for instance, apparently do not exercise enough of this native endowment to be of

much value in judging the general intelligence of an individual. Hence they are being discarded as measures of intelligence. The same may be said of most tests functioning on a perceptual level.

Fortunately, after a period of about twenty years of work in attempting to measure intelligence, we have a symposium by a group of thirteen leading American psychologists on what intelligence is, how it may best be measured, and what the next steps are in this field.¹⁶

Terman in attempting to answer these questions in the symposium has tersely stated the situation in regard to the value of a right conception of general intelligence. After endorsing the conception of Meumann that we should first find out "what is demanded of intelligence and then analyze the mental functions which meet that demand," he says:

If we accept this view it is evident that the important intellectual differences among men will not be found on the sensory, perceptual, or purely reproductive level. It is well known that a moron may be able to see, hear, taste, or smell, react to a signal, balance a bicycle, steer an automobile, or cancel A's about as well as an intellectual genius. The latter would be somewhat his superior in memory for non-sense syllables, would excel him more in logical memory, and would outclass him hopelessly in the ability to distill meanings from the raw products of sensation and memory. The essential difference, therefore, is in the capacity to form concepts to relate in diverse ways, and to grasp their significance. *An individual is intelligent in proportion as he is able to carry on abstract thinking.*

In answer to the retort that some may make that he is simply singling out a particular mental trait for special worship, that other traits are just as valuable as abstract thinking, and that it is a kind of intellectual snobbery,

¹⁶ "Intelligence and Its Measurements: A Symposium," *Journal of Educational Psychology*, Vol. 12, March and April, 1921.

that holds that general intelligence is manifested chiefly by one's ability to do abstract thinking, he says:¹⁷

Civilization with its science, art, government, religion, philosophy, and systems of credit, is unthinkable except as a product of concept elaboration and symbolic thinking. . . . It cannot be disputed . . . that in the long run it is the races which excel in abstract thinking that eat while others starve, survive epidemics, master new continents, conquer time and space, and substitute religion for magic, science for taboos, and justice for revenge. The races that excel in conceptual thinking could, if they wished, quickly exterminate or enslave all the races notably their inferiors in this respect. Any given society is ruled, led, or at least molded by the five or ten per cent of its members whose behavior is governed by ideas. The typical pick-and-shovel man does his thinking chiefly on sensori-motor and perceptual levels. Add a little more ability to think on the representative level and he may be able to repair your automobile, build you a house according to an architect's specifications, or nurse you in illness. Add a large measure of ability to associate abstract ideas into complex systems and he can design a new type of engine, draft the plans for a skyscraper, or discover a curative serum.

In regard to the next step in research in the measurement of intelligence, the general consensus of opinion of the men taking part in this symposium was that the immediate task is to refine the tests and catalogue the characteristics that are recognized as belonging to higher intelligence; that is, those elements that emphasize, more than any now existing, deliberation and sustained rational ability. There must be a courageous attack upon the problem of measurement of other than intellectual factors. The problem of special aptitudes must be attacked and the general technique of measurement improved. While this programme is a broad one, nevertheless, the problems are being vigorously attacked and progress is being made.

¹⁷ *Ibid.*, pp. 127-128.

One of the fields that is receiving special attention at the present time is the measurement of the non-intellectual traits of individuals. The work of Dr. Downey on the "Will-Profile" suggests the possibility of supplementing our intelligence examinations by objective measures of the so-called character-traits.

Dr. Haggerty in summing up the work of mental measurements for the past year, says: "If a single summary phrase were useful to indicate the drift of current discussion, we might choose 'the inadequacy of intelligence' as a suitable title."¹⁸

It is obvious to the careful experimenter that tests of the type now in use do not give all the information that we need to know about children. Either the definition of intelligence must be broadened to include other traits, or we must conclude that there are traits of a non-intellectual type that determine to a large degree the success or failure of an individual in life. The functioning of the eye muscles may, for instance, determine the amount of reading an individual can do, and thus condition his entire life programme.

Industry, which is apparently beyond the limits of what we call intelligence, has much to do with success. Haggerty¹⁹ directed an experiment a few years ago, the purpose of which was to study the characteristics of 50 men admittedly successful and 50 others who were obviously failures in life.

When the data were combined for each of the successful men the result showed clearly that, in the combined opinions of all the judges, the quality most apparently conducive to success was *industry*, which was defined as

¹⁸ "Recent Developments in Measuring Human Capacities," *Journal of Educational Research*, Vol. 3, April, 1921. Address delivered before the National Association of Directors of Educational Research at Atlantic City, N. J., March 3, 1921.

¹⁹ *Ibid.*, p. 245.

"thorough, persistent, painstaking, enduring" and the opposite of "lazy, sluggish, indifferent, superficial." The nine traits ranking next in order were: "efficiency, attentiveness, loyalty, prudence, honesty, adaptability, sympathy, tactfulness, and cheerfulness."

The most of these traits are beyond the limits of what we ordinarily call general intelligence. Such non-intellectual traits as industry, loyalty, honesty, tactfulness, sympathy, and cheerfulness weigh heavily in favor of success, and such other non-intelligent traits as self-assertion, pride, conceit, jealousy, quarrelsomeness, suggestibility, and intolerance make their contribution in the direction of failure. Either our tests of intelligence are inadequate or intelligence itself is inadequate to produce success.

We have many cases of children in the public school with a high I.Q. who do very poor work in school, and also many cases with a normal I. Q., or even an I. Q. below normal, who make the best grades in the class. The latter are invariably industrious and persistent. Haggerty reports the case of a boy who stood third in a class of 60 in a series of intelligence tests which included the following: opposites, analogies, hard directions, verb-object, Trabue completion, and the Thorndike Reading Scale, Alpha 2. He was later examined with the army examination A, in which he scored 325 points placing him easily in the upper 10 per cent of high school freshmen and the equal of many college students. He scored 179 points on the Otis Scale; yet during his four years in high school only twice did he achieve a mark as high as C on a five-point scale of marks.

In contrast with this student was a girl in the same class whose score on the army examination A was 231, who scored average on the initial tests (ranking 23 in a group of 60 entering pupils) and whose I.Q. (Terman) was 108.

In her four years of high-school work, only five times did this girl make a grade as low as B in an academic subject. Twenty-eight times out of 33 her marks were A, and she was classed by her instructors as the best student in the class.

Examples like those cited by Haggerty can be found in all large school systems.

Pressey notes that "if we wish to foretell success in school we must obtain a measure of school attitude."

Terman insists that "mental tests should be supplemented by ratings on character traits and by educational tests."

Haggerty points out that "it is not at all probable that a perfect measure of intelligence would give a perfect correlation with school success or with success in later life. A more accurate measure of intelligence would only render the inadequacy of intelligence more apparent for the simple reason that success is not quantitatively coterminous with intelligence but with intelligence in combination with other significant human traits not subject to evaluation by tests of the type currently used as measures of intelligence."²⁰

In the "Will-Profile" mentioned above, Dr. June Downey has attempted to design a scale that consists of 12 objective tests designed to measure such personal qualities as assurance, flexibility, speed of movement, motor impulsion, resistance, tenacity, coördination of impulses, freedom from inertia, motor inhibitions, care for detail, speed of decision, etc.

The author claims that these tests have considerable general characterological significance and that they can be used to advantage in getting the general temperamental pattern of an individual and they may also determine specific combinations of traits, and, in conjunction with

²⁰ *Ibid.*, pp. 246-247.

intelligence tests, afford in many situations a basis for conservative prophecy.

Healy points out another phase of the measurement of intelligence that is worth noting. There is a type of individual he calls "verbalists" which is characterized by an ability to handle language above his ability along other lines.²¹ "On account of the ability of this type to handle language well, the members of this group are not properly placed by the ordinary tests of social intercourse. The common method of passing judgment on people is, of course, through conversation and also questions, and if one gets answers that follow properly, that are consequential and coherent, why then without more ado one infers the answerer to be practically normal."

It is interesting to note that the city superintendents and high-school principals of the Council of Administration in the State of Kansas endorsed a plan on January 20, 1921, whereby many of these non-intellectual traits and traits indicated by Dr. Downey's "Will-Profile" are to be taken into consideration in the awarding of grades in the high schools in the State of Kansas.

The definition of Grade A as endorsed by that Council is given below:

Grade of A

1. *Scholarship.* Exceeding expectations of instructor.
2. *Initiative.* Contributions exceeding the assignment.
3. *Attitude.* Positive benefit to class.
4. *Coöperation.* Forwarding all groups of activities.
5. *Individual improvement.* Actual and noticeable.

Educators are beginning to ask which is the more significant inquiry concerning a candidate for admission to college that has just completed a four-year high-school

²¹ W. O. Healy, *The Individual Delinquent* (Little, Brown, & Co., 1915), pp. 473 et. seq.

course: "What subjects has he previously studied?" or "What is his general intelligence or ability?" The subjects a high-school student has previously studied are significant primarily because they may give a measure of his general ability, rather than because they indicate a knowledge of any particular fact. The intelligence examination can, therefore, in a general way, be valuable in connection with the selection of students for college admission only as supplementing the high-school record and not as a substitute for it.

The question naturally arises as to how the regular room teacher may utilize these tests for more efficient school work. We can speak quite dogmatically and say that they are a great improvement over the personal judgment of the teachers.

A few suggestions will be offered here relative to the use of these tests by the regular room teacher.

1. The teachers should familiarize themselves with some of the standardized intelligence tests now in use.

2. Care should be taken to see what these tests are designed to test. A familiarity of the various traits of the mind gained through the careful study of the tests will quicken the teachers to detect particular traits very early in the child's career.

3. If the children are given mental tests, the teacher learns the type of minds with which she has to work.

4. Reasonable improvement can be determined only after the mental capacity of the individual child has been ascertained. If progress in school achievements has been slow, the teacher has a good defense if the general intelligence is low. On the other hand, if the children have high intelligence quotients and progress has been slow, the burden of the proof that the school was well taught would lie on the teacher. Of course, there are other factors that enter into the problem that must be taken into consideration.

5. *If regular room teachers, not specially trained to give intelligence tests, should give such tests, the directions for giving the tests must be followed to the letter and conclusions and implications must be conservatively drawn.*

THE MEASUREMENT OF INTELLIGENCE 145

6. Don't treat tests as "educational curiosities." Their purpose is to give information about the developing mind. They suggest educational policies. They indicate a more scientific selection of subject matter.

7. If we are to have better tools for measuring intelligence, we must discover through the use of the tools we now have wherein they are deficient.

8. The spiritual admonition of the Apostle Paul when he said, "Prove all things; hold fast to that which is good," was never more needed in the field of religion than in education.

9. Every teacher should learn the principles involved in test making and in giving tests, not always for purposes of diagnosis of children but for the guidance of her own behavior towards them.

10. Psychological tests will enable us to predict with a fair measure of scientific accuracy the extent to which pupils will avail themselves of the opportunities which are set before them. They do not constitute a psychological clinic; nevertheless, they are valuable in diagnosis. Their purpose is to make a mental analysis of the child in order to discover the assets and defects so that the examiner may prescribe the proper treatment.

BIBLIOGRAPHY

1. BURT, C., *The Distribution and Relations of Educational Abilities* (King and Son, London).

2. COURTIS, S. A., *The Gary Public Schools; Measurement of Classroom Products* (General Education Board, New York, 1919).

3. HAGGERTY, M. E., *The Intelligence Examination* (World Book Co.).

4. HEALY, W. O., *The Individual Delinquent* (Little, Brown & Co., 1915).

5. HAGGERTY, M. E., "Recent Developments in Measuring Human Capacities," *Journal of Educational Research*, Vol. 3, April, 1921.

6. HOLLINGWORTH, LETA S., *Vocational Psychology* (D. Appleton and Co., 1916).

7. "Intelligence and its Measurements; A Symposium," *Journal of Educational Psychology*, Vol. 12, March and April, 1921.

8. JUDD, CHARLES H., *Measuring the Work of the Public*

Schools, Cleveland Educational Survey (Russell Sage Foundation, New York, 1916).

9. LINK, H. C., *Employment Psychology* (The Macmillan Co., 1916).

10. MÜNSTERBERG, HUGO, *Psychology and Industrial Efficiency* (Houghton Mifflin Co., 1913).

11. *National Intelligence Tests*, prepared by Haggerty, Terman, Thorndike, Whipple, and Yerkes (World Book Co.).

12. National Society for the Study of Education, the various *Yearbooks* (Public School Publishing Co., Bloomington, Ill.).

13. *Otis Group Intelligence Scale*, designed by Dr. Arthur S. Otis (World Book Co.).

14. PINTNER, RUDOLF, and ANDERSON, MARGARET M., *The Picture Completion Test*.

15. PINTNER, RUDOLF, and PATERSON, DONALD, *A Scale of Performance Tests* (Warwick and York, 1917).

16. ROSSOLIMO, "Mental Profiles; A Quantitative Method of Expressing Psychological Processes in Normal and Pathological Cases," *Journal of Experimental Pedagogy*, Vol. 1.

17. RUSK, ROBERT R., *Experimental Education* (Longmans, Green & Co., 1919).

18. TERMAN, LEWIS M., *The Measurement of Intelligence* (Houghton Mifflin Co., 1916).

19. *Terman Group Tests of Mental Ability*, designed by Lewis M. Terman (World Book Co.).

20. YERKES, BRIDGES and HARDWICK, *A Point Scale for Measuring Mental Ability* (Warwick and York, 1915).

CHAPTER V

THE NEED FOR DEFINITE MEASUREMENTS OF SCHOOL ACHIEVEMENTS

Time Consumed in Giving Examinations.—It is estimated that on an average each teacher gives as many as twenty examinations each year and that it takes approximately three hours to give each examination and to grade the papers.¹ If we estimate that there are 600,000 teachers in the United States, this would mean that 36,000,000 hours are spent in giving examinations. Therefore, if the time can be lessened and the accuracy of the measures increased, the effort is well worth while.

That the examination system is here to stay is an obvious fact. The need for some method of determining the efficiency of the educational processes is so obvious that no argument is needed to substantiate it. The folly of putting children through educational processes day after day and never knowing what the results are can be endorsed neither as a principle nor as a matter of expediency. Schools cannot be efficiently operated without a system of examinations any more than a business man can run his business without invoicing his stock occasionally. The best known and, in fact, the only way to determine the value and status of an educational process is to take an invoice of the products. typ

One cannot work long in the field of experimental education without coming to the following conclusions: (1) *Examinations of some kind are necessary.* (2) *The exami-*

¹ "A New Kind of School Examination," *Journal of Educational Research*, Vol. 1, pp. 33-46.

nations must be of such a nature that they may be given by the regular room teacher. Only rarely can they be administered by the superintendent in person or by some one acting for him.

Though convinced theoretically, many administrators find it hard to see just how measurements can be effectively carried on to advantage in their schools. The accusation is made that examiners testing their schools carry away the results and, if they ever get back to the schools, they are many times so intangible that they mean nothing as a factor in determining a school policy.

We must build our ideal system of education synthetically, taking the best methods from each of the prevalent groups of theories. But we cannot determine best methods until we measure our processes. People generally agree in measuring a product, if they can agree on the measuring stick.

Measurement in any department of natural science is the comparing of a given magnitude with some convenient unit of the same kind, and the determination of how many times the unit is contained in the magnitude. The unit of measurement is conventional. Its choice is simply a matter of practical convenience.

Unfortunately, until quite recently, there have been no objective units of measurements of general acceptance in the field of education. A good teacher, or a good student, or a good school was purely a matter of judgment with no commonly accepted measuring device to verify the judgment made.

Attitude of Teachers and Pupils Toward Examinations.

—Generally speaking, most teachers and pupils dislike examinations. Since this feeling is so general and since so much has been said derogatory of examination, the subject challenges us to a careful study as to the causes of this state of affairs. Are examinations in themselves just

naturally repulsive and obnoxious to pupils and teachers? Do pupils dislike to have their educational achievements measured? Each of these questions must be answered in the negative. There is no part of the educational process that pupils like better than to have an unbiased statement as to their school achievements. There is nothing fundamentally distasteful about an examination either for the teacher or pupil. Yet the examinations as now given are not satisfactory. Since they are necessary and are not intrinsically and fundamentally repulsive, the cause of their dislike must be referred to one or more of the following: (1) the methods of devising the examination; (2) the methods of administering the examination; or (3) the methods of scoring the papers. Improvements along one or all of these lines will help to remove the obnoxious features from examinations and make them a pleasure instead of a piece of drudgery that must be tolerated.

School achievement tests, discussed in this chapter and Chapters VI and VII are nothing more than improved examinations. It is confidently believed by the writer that improved methods of devising and administering examinations and scoring examination papers, such as we shall have eventually in the form of standard tests, will make contests in school achievements as interesting as athletic contests are at the present time. There is no doubt but that the case of scientific measurements has been argued and won. Of course, the fundamental thing in school achievement tests is not to provide mere entertainment but to devise a system of measurements that really measures. Nevertheless, if the general attitude of teachers and pupils regarding examinations can be changed so that they may be looked upon as a pleasure instead of drudgery as they are considered now, attempts at improving them are worth while.

That both teachers and pupils need "professional

hurdles," quantitatively stated goals, set before them, against which their developing effectiveness may be frequently checked up, is a recognized principle in education. It is one method of keeping them up to optimum efficiency.

The Marking System now in Vogue.—If there is any one phase of school work in which both the teachers and the public apparently have had an abiding faith, it has been in the ability of teachers to express definitely and concisely in per cent, letters, or adjectives, the exact progress a pupil has made in his school work. As evidence of this faith we have allowed these marks to determine the fitness of candidates for college; their eligibility for athletics; scholarships and fellowships, which amount to hundreds of thousands of dollars every year, have been granted with practically no other evidence of the candidate's fitness; these marks have determined the fitness of candidates for civil service positions, admission to Phi Beta Kappa, "*cum lauda*," "*magna cum lauda*," and other special honors granted certain members of graduating classes who were successful in getting the requisite number of A's or H's or other high marks of distinction.

Parents look forward with a great deal of anxiety to the time when Johnny will bring home, in the form of a monthly or quarterly report card, a record of his achievements. A fair "sprinkling" of A's and B's is sufficient evidence to convince the average father or mother that Johnny's progress is satisfactory, so with a great deal of pride, but little real information, the report card is signed and returned to the teacher. If the grades were poor the parents might question the honesty of the teacher but never the reliability of the marking system.

Grades A, B, C, and D usually represent degrees of excellence between 100 per cent and 60 per cent of "something"; no one knows exactly what. A pupil's "passing mark" is usually a grade equivalent to 75 per cent of

"something," and if he gets an average grade of 95 per cent of this "something," he may be valedictorian of his class.

If school marks are so inefficient, the question immediately arises as to whether some "inventive genius" or "wizard" has invented some system of scales and units by which school achievements may be measured with the exactness that the apothecary compounds his medicines, or the mechanic measures the diameter of a piston head.

The answer to this question is perfectly definite. *There are no scales or units yet invented that will measure school achievements with anything like the exactness that it is possible to get in the physical sciences.* It should be noted, however, that the precision that the natural sciences enjoy was not developed in a day. The sciences have been evolved a step at a time from humble beginnings to the high state of perfection they now enjoy. The Wright brothers did not equip their first aeroplane with a Liberty motor. Robert Fulton's steamboat would be a sorry-looking spectacle beside the great floating palaces that now cross the Atlantic in three or four days; and Stevenson's wheezing little locomotive would certainly be at a disadvantage in competition with a "twentieth century limited." Each of these great inventions has had the most humble beginnings, and their evolution may be traced a step at a time until the present state of efficiency is reached.

In the same way we shall have to refine our standards and units of educational measurements. There has been too much absolutism in education and too little of a realism that sees the good and bad in all and diminishes the bad and augments the good. If we adopt this view we become really empirical, living through each educational experiment to incorporate it into a growing treasury of tested theory, not deducing success or failure from metaphysical or doctrinaire prejudice.

It may not be too much to predict that twenty-five years from now our present scales and standards will be as much out of date as Fulton's steamboat. Something like perfection will have been reached when we are able to construct scales by scientific methods that can be applied as the foot rule is now applied, regardless of time, or place, or person.

Scientific Measurement of School Achievements Is New.

—It is only within the last dozen years that any real progress has been made in the scientific measurement of school products. In 1908 Dr. C. W. Stone, a student under Thorndike, published his arithmetic test, and the following year Thorndike presented his scale for handwriting before the American Association for the Advancement of Science, then meeting in Boston. These two tests represent the real beginnings of the scientific measurement of educational products. Thorndike has been called the father of the educational measurement movement in America, and Dr. J. M. Rice, mentioned in the introductory chapter, the grandfather. Inspired by the work of Rice, Stone, and Thorndike, Curtis soon followed with an arithmetic test in the four fundamentals which went through a number of revisions until it took the present form now known to us as The Curtis Standard Research Tests in Arithmetic, Series B. At the present time the number of standard educational tests that have been developed for the various subjects is probably somewhere between 100 and 150.

In spite of the fact that practically every educational meeting of note for the last ten years has devoted a great deal of the time to the subject of the scientific measurement of school achievements, and also the fact that educational literature is replete with discussions of the new movement, yet at the present time most teachers accept the fact of educational measurements in a *passive* way,

and a large majority have never made use of these new educational tools.

On the other hand, there is a minority group of teachers who have looked upon these tests and scales as instruments for refining their cruder processes, and they are making rapid strides in this direction. The fault, however, does not lie altogether with the teachers. The administrative machinery, both from the standpoint of the state and local district, is organized and operated on the old system. Grades are still recorded and promotions made on the old basis. An administrative change must be brought about before the new movement has the complete right of way.

Those in the Profession Must Take the Initiative for Improvement.—Professional improvement must come from *within* and not from pressure without the profession. The medical profession has not advanced because the lawyer, or the minister, or the business man forced up their standards. They were forced up by the leaders in the profession. School management will never become more scientific than the teachers themselves make it. The more teachers that may be induced to accept improved methods, the sooner the teaching profession will reach a stage where teachers may be held morally, if not criminally, liable for malpractice just as they do in the medical profession. Before this can be done teachers must be thoroughly convinced that the old methods are inadequate and that the new ones offer better opportunities so that they will not hesitate to adopt the new. Of course we cannot expect a complete transition over night. Changes have not been wrought in other sciences that way, and progress is pretty much the same in all sciences. The "arts of healing," for instance, of our ancestors, the product of ages of selective effort, have given way to the modern science of medicine. However, some people still cling to the dogmas and cures of the older medicine as to cherished heirlooms.

Since the transition from magic and faith cures, the medical profession has become so strongly entrenched by the accumulation of scientific facts that it yields to no social force. Even the military takes orders from these men of science.

There are, of course, new fields being explored where it is still a matter of opinion as to what procedure is best; but along with these there is a procedure in some of the more familiar fields that is so definite that any departure from it makes the practitioner liable to a fine for malpractice. A more profound insight into scientific methods and a greater sensitiveness of our shortcomings are the twin forces that are operative in convincing men that human incapacity, suffering, and waste can be reduced and life made better through more purposeful use of scientific knowledge as it becomes more accessible.

Purposes of Educational Tests Are Not Generally Understood.—The work in educational reform has been hindered a great deal by a lack of understanding as to what the purposes of tests are. Monroe cites the case of a school superintendent in a city of more than 50,000 population stating that he did not believe tests had much merit. His reasons were that he had given the Courtis Standard Research Tests in Arithmetic and the children did not do any better after the tests were given than they did before. Just as though standard tests were teaching devices that would improve their arithmetic ability. That is analogous to the case of a mother who, being anxious for her baby to gain in weight, said that she would not weigh her baby any more because weighing it from time to time did not increase its weight.

The Problem to Be Solved.—The first problem that confronts one attempting to measure school achievements is that of finding some convenient units of measure to apply to the achievement in question. If one wants to know

the length of a board or the width of the street, there are several convenient units with which to measure them. They may be measured in feet, inches, yards, meters, centimeters, or other derived units. If one wants to know how heavy a thing is, there are well-defined units at his disposal such as the ounce, pound, gram, and kilogram. Temperature is measured in degrees, electricity in kilowatt-hours, etc.

When one attempts to measure the handwriting of an individual, or how well he can draw or spell or how well he can read and cipher, there are no such convenient units at his disposal. One of the first problems to be solved, therefore, by those desiring to measure school achievements is to devise scales and units that may be used in measuring these products. In linear measure and measures of weight, the steps on the scale are the same, that is, the difference between 1 pound and 2 pounds is the same as the difference between 21 pounds and 22 pounds, or between 48 and 49 pounds. The question immediately rises as to whether scales for measuring educational products may be made and used as the above-mentioned scales are used in other sciences.

What School Achievement Tests Measure.—The first thing to be noted is what a test measures. Professor Curtis has given us three terms which are helpful in bearing this thought in mind. These terms are: (1) *capacity*, or the original intellectual endowment of the educand when he enters school; (2) *ability*, which is capacity plus training; and (3) *performance*, or that which the individual actually does on a given test. It is a rare case when one's ability is equal to his capacity. That is, the native endowment is such that, generally speaking, it might have been cultivated into more ability. It is also true that performance is usually less than ability. It is a rare case when one is taking a test that all the ability is utilized. When

an eighth-grade boy takes a test in addition, for instance, and works as hard and rapidly as he thinks it is possible, he could, no doubt, add another problem to his credit if the reward offered were made large enough. In school achievement tests, we are testing performance and not capacity or ability. It should be noted that the word "performance" is used here in a different sense from that used in discussing tests of intelligence. There it was used in a somewhat restricted sense, meaning activities which do not depend on the ability to read.

Experimental Evidence to Show that School Marks Are Inadequate.—As was indicated in the introductory chapter, one of the first things that must be done to make measurements of educational products more nearly exact is to convince teachers that the old marks now being used are inadequate. It seems fair to assume that with an efficient measuring device or yardstick any number of competent persons measuring the same thing ought to get approximately the same results. It is not expected that the results will be exactly the same because exact measurements are impossible in any field. If, for instance, 116 competent persons were given a measuring stick and asked to measure the length of a board the length of which was known to be somewhere between 0 and 100 inches, and, if the results obtained varied from 28 to 92 inches, we would conclude that there was something wrong with the measuring device. If the measuring were done with a foot rule divided into inches, a variation of an inch or two would be the maximum we would expect.

Starch and Elliott investigated the accuracy with which 116 teachers were able to measure the merits of a geometry examination paper. A facsimile reproduction of the examination paper was made and sent to each of the high schools included in the North Central Association of Colleges and Secondary Schools with the request that the

geometry teachers in the various schools mark it on a basis of 100 per cent. One hundred and sixteen teachers, in as many schools, complied with the request. The grades ranged from 28 to 92 per cent. The conditions are analogous to those mentioned above in measuring the board. The supposition is that the teachers asked to measure the merits of the paper are competent persons or they would not have been employed to teach mathematics in the high schools. Another factor that deserves special attention is the fact that a geometry examination paper is supposed

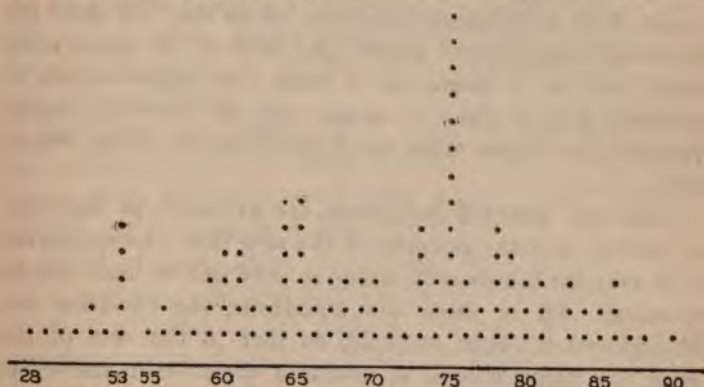


FIGURE II. DISTRIBUTION OF MARKS ASSIGNED TO ONE GEOMETRY PAPER BY 116 TEACHERS

to be of such a nature that it may be graded with a much higher degree of accuracy than an examination in literature, history, or reading, for instance. Figure II shows the distribution of marks as they were compiled from the reports of this test. Two of the marks were above 90 while one was below 30. Thirteen teachers gave the paper a grade of 75 per cent. Twenty-seven marks were above 80 while 20 others were below 60. Considering a grade of 75 per cent as a "passing mark," 47 of the 116 teachers considered the paper to have sufficient merit "to pass," while

69 thought the paper not worthy of a passing mark. This gives a range of more than 60 per cent. Two or three inches would be the maximum variation we would expect in the case of measuring the board above mentioned.

The case of measuring the board would have been more nearly analogous to that of measuring the examination paper if instead of using a rigid foot rule we had used a piece of elastic. In that case it would all depend on the tension of the elastic as to what results were obtained. It is as difficult even for the same teacher to grade a paper twice, with a sufficient time interval so that she does not remember her former grade, and give it the same score each time, as it would be to make two measurements of a board with a piece of elastic and get identical results because she forgets what the former tension of the elastic was.

Teachers' unaided judgments are as elastic as the piece of rubber, and the purpose of the new tests and standards is to give the marks some fixity so they will be more nearly constant. If, by these new standards, the variation can be reduced by even one-half, so that in the case of the geometry paper instead of having a variation in the scores of more than 60 per cent (from 28 to 92,) it may be reduced to 30 per cent, or even less, a wonderful change for the better will have been wrought.

Starch and Elliott not only investigated the ability of geometry teachers to mark geometry papers but made similar tests in history and English. In all three cases the variations were very great. The papers in each case were presumed graded by experts in the various fields.

Marking System Inefficient because It Does Not Indicate Progressive Degrees of Merit.—If one were asked to grade a composition, one of the first things he would want to know would be what age, grade, or class of pupil wrote it. If told it was written by a pupil in the second grade,

a boy six or seven years old, it might get a mark of 95 or 100 per cent, but if told that it was written by an eighth-grade pupil, a mark of 60 per cent might be given. Now it is evident that the composition has the same merit whether written by a second-grade pupil or a college graduate. Instead of having one scale for grading compositions the person grading the paper uses a different scale for every grade in school. A boy may get 95 per cent in his composition work from the first grade to the time he graduates from college. Since it is obvious that he can write a better composition in college than he can when he is only a second-grade pupil, there should be some way of expressing in definite units of accomplishment the advancement made.

Definition of a Scale.—It is to correct this defect that educational scales are being devised. The word “scale” comes from the Latin word *scala*, which literally means a ladder, or a staircase, a long straight article divided into a series of equal steps and readily lending itself to use as a measure. In the physical sciences the scale reaches from the *lowest* to the *highest*. A scale for measuring weight reaches from the lowest, or smallest amount we desire to measure, to the highest, or greatest. A scale for measuring length or time reaches from zero to some upper limit as high as we desire to go. As used in education it is thought of as a linear rule extending from the worst to the best; or from an educational product of least merit to one of greatest merit; from problems of least difficulty to those of greatest difficulty, and so on. Measurement is in terms of relative merit, showing whether a given product or performance represents an achievement that is halfway along the scale from worst to the best or at a point 80 per cent of the distance from the zero point to the high end, or some other definite location.

We are apt to ignore the relative worth of things. When

we attack a standard in school, we set absolute perfection as our own standard. We fail to consider the fact that the final increment of any subject is frequently obtained at an expenditure of energy out of all proportion to its worth. Scales will help in discovering what it costs in time, energy, and practice to reach the various standards set. We shall expect that curricula may be constructed on the basis of aims or objectives that are scientifically determined and are not the chance product of personal preference and opinion. The function of the scale is to take the score resulting from the test and interpret it in terms of relative merit.

The test is analogous in many respects to the pea, or sliding weight, on the scale beam and is used to determine the exact location of the individual on the scale. Making a test too easy would not determine the position of an individual on the scale any more than setting the pea at 10 pounds would determine the weight of a body weighing approximately 100 pounds.

Just as one estimates the weight of an article by its size or by lifting it and sets the pea at the estimated weight on the scale beam, so the test must be arranged so that the point reached on the educational scale will lie within the range of the test. If, for instance, all the problems are solved in a speed test in arithmetic before time is called, one would not know whether the upper limit of a child's ability had been reached any more than he would know the weight of a thing if he were to set the pea at 100 pounds on the scale beam and did not move it above or below that mark to see if the article would weigh more or less than 100 pounds.

Progress from grade to grade cannot be measured accurately if the per cent system, as now used, is to be employed. The Thorndike Handwriting Scale, while inefficient in many ways, possesses the essential character-

istic of indicating degrees of merit from the lowest to almost perfection. The various specimens of handwriting start with Quality 4, which is recognized as handwriting but almost entirely illegible, and go to Quality 18, which is well on the way towards perfection. The quality of a specimen of handwriting is determined by sliding the specimen to be measured along the scale until that quality is reached which most resembles the sample in question. The age or grade of the writer has nothing to do with the quality. An eighth-grade boy and a first-grade boy may write specimens of exactly the same quality.

An Ideal Scale Must Have Equal Steps and Each Step Must Bear a Definite Relation to the Zero Point.—It was indicated in the introduction that our present marking system follows no known rules of mathematics. A grade of 80 per cent does not mean that the product is twice as good as that given a grade of 40 per cent, and 75 per cent is not one and a half times as good as a grade of 50 per cent, because there is no zero point established to which these marks bear a definite relation. Our new educational scales are built on the same plan as the scales for length and weight in the physical sciences. In order to obtain this characteristic of having successive steps on the scale indicate progressive values increasing by equal amounts, the new scales are based on the theory of the so-called *normal distribution of intellectual ability*. This distribution is represented graphically by a bell-shaped curve, or it is like the cross-section of a pile of sand dumped from a cart. The following illustrations of a normal distribution surface will help clarify our conception as to how the new educational scales are made.

If one were to pluck all the leaves from an oak tree and arrange them in a long line according to their lengths so that the shortest leaves would be at the left end of the line and the longest ones at the right, he would find that

he had very few exceptionally short leaves, a great number of medium length, and very few exceptionally long ones. If these were represented by a surface of frequency, the diagram would be the bell-shaped curve mentioned above (see Fig. I, page 77). Now, if the number of leaves of various lengths were represented by the height of the curve above the base line, at the extreme left the curve would come very close to the base line because there are very few leaves exceptionally short. As the leaves grow progressively longer their number also increases, and the curve rises rapidly from the base line, first concave then convex, until the peak is reached. It then descends toward the base line making a symmetrical curve.

If 10,000 adult males, all belonging to the same race, were chosen at random and lined up in the order of their heights, the short ones at the left end of the line and the tall ones at the right, their distribution would approximate very closely that mentioned above in reference to the leaves. That is, there would be a few exceptionally short ones, a great many of medium height, and very few giants. The weights of individuals, the strength of their grip, and, in fact, most of their physical traits are found to obey the same law.

The question then arises as to whether the intellectual traits of individuals distribute themselves in the same way. It has been found that if tests are given in any grade as the fifth, for instance, in the subject of spelling, a few exceptionally poor spellers are found, a large group of spellers with median ability, and a very small group of exceptionally good spellers. The same distribution is found if any considerable number of children in any grade are tested in any subject. In other words, the same law that operates in the physical and biological sciences is also operative in reference to intellectual traits.

The problem of scientific scale building is further sim-

plified by the fact that mathematical laws concerning the characteristics of the surface of normal distribution have been most accurately determined, and by the application of those laws it is possible to locate and determine the steps on the different scales for school achievement.

Now since it is found that intellectual traits distribute themselves according to a normal or symmetrical distri-

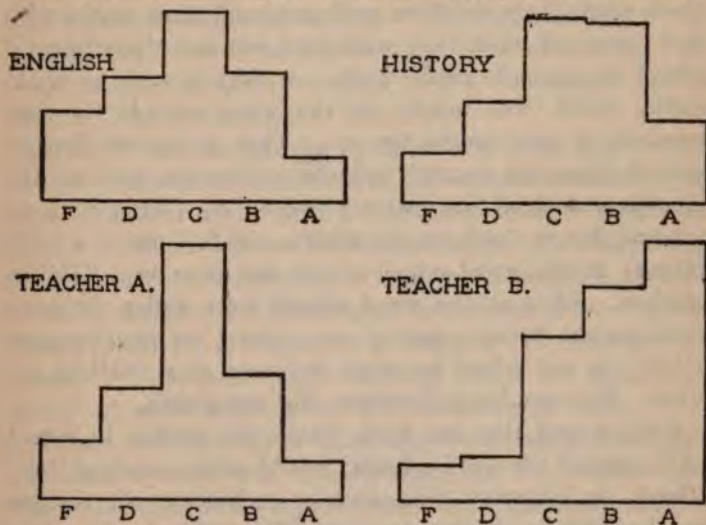


FIGURE III. ABOVE, DISTRIBUTION OF MARKS IN ENGLISH AND HISTORY IN THE UNIVERSITY HIGH SCHOOL OF THE UNIVERSITY OF CHICAGO. BELOW, DISTRIBUTION OF MARKS OF TWO TEACHERS IN THE SAME DEPARTMENT (after Johnson).

bution, if one were to examine the grades given by teachers and found they were distributed in some other way than according to a normal distribution of frequency, it would be perfectly legitimate to question the marking system.

An investigation of this kind was made by F. W. Johnson, principal of the University High School of the University of Chicago of the grades given for the school years 1907-8 and 1908-9. Figure III shows the distributions

of the grades.² It may be seen that the distribution of grades here do not conform to a normal distribution surface.

Another investigation which throws some light on the inadequacy of school marks is that made by Dr. F. J. Kelly. In 1913, Dr. Kelly made an investigation in four ward schools in Hackensack, N. J. to determine the marks given sixth-grade children, and compared these marks with those received when they went to a common departmental school for seventh grade work. A subject such as arithmetic, which was taught in the ward schools by four teachers, is now taught by one. That is, by the departmental plan one teacher had the arithmetic, another the language, a third the history, and so on. This gave an opportunity to check up the grades and find out if a "G" (Good) in one ward school meant the same as a "G" in another. Since all the ward schools were under the same management, being a part of one system, we would expect a "G" in one school to mean the same as a "G" in another. This condition, however, did not obtain.

Kelly found that for work which the teacher in school "C" (one of the ward schools) would give a mark of "G" (Good) in language, penmanship, or history, the teacher in school "D" (another ward school) would give less than a mark "F" (fair).³

More Exact Measurements Will Make Education a Science.—The application of scientific measurements to school products is doing more to make education a science than any other contributing cause. It is giving education tools with which to work. In this field we are making wonderful strides. The soil is virgin; hence it will take educators a long time to put it on a plane with the other

² *School Review*, Vol. 19, pp. 13-24.

³ F. J. Kelly, *Teachers' Marks*, Teachers College Contribution to Education, No. 66, 1913, p. 7.

sciences. For instance, a few years ago a reading test seemed impossible; to-day, we have mastered the distinction between oral and silent reading. We have good methods of measuring some of the more common types of deficiencies, and we know the rate of progress which is normal in the more obvious phases of interpretation. The advantages of tests are in the same line as the advantages of the thermometer over saying, "The heat is stifling," "It is very hot," "It is boiling hot," etc.; or of saying, "He is the tallest person I ever saw," or "the biggest in the state of Massachusetts."

It may not be too much to predict that some day concentration of attention, ability to attack various kinds of problems, clearness of insight, power of inference in various fields, and other abilities will be measured. An educational product, as a composition, is usually a complex, and its measurement is more like measuring a house or an elephant than measuring a length or a volume. It must not be expected, therefore, that a thing may be completely described quantitatively when it is measured once.

Supervision Improved as Ability to Measure Increases.
—Standard tests will greatly improve our supervision. Our children will be more rapidly and effectually taught. Inefficient teachers will no longer be able to hide behind our ignorance. Dubious aims of education and aims too remote to be effective to the general practitioner will be replaced by goals that are in sight, by a motive which is dynamic and energizing, and by an appeal which will spur pupils on to a greater effort. The pupil wants to know how he is getting along. He wants to know where he stands in reference to an impartial standard. He wants to know if his performance each succeeding day and week brings him a little nearer the coveted goal. The teacher is eager to know what degree of success her efforts have won as measured by the quality of the work done by her

pupils. It must be borne in mind, of course, that an educational product is a complex. It is the resultant of a great many causes, of which the school is but one. The home, the native capacity, the amount of study done by the pupil, the influence of the street, the playground, the state of health of the educand, and many other factors enter into this complex. To isolate and measure the school factor is a difficult task to perform. One must be very careful about drawing conclusions too hastily when schools are compared. Suppose for instance, school A were compared with school B in the four fundamentals of arithmetic and the median score for school A was three problems more than school B. Teachers must not draw the conclusion that the difference was caused exclusively by better teaching in school A. It may be that more time was devoted to the four fundamentals in arithmetic in school A than school B. Or the native capacity of children in the former school may have been much better than that in the latter; they may have studied harder; they may have received more help and encouragement from home; the general status of health may have been better than in school B. Along with these factors may have been the fact that school A had better teachers and better teaching methods than school B. Or, the opposite may have been true. Teaching methods cannot be compared in this way. If one wanted to compare the merits of two methods of teaching a certain subject, as, for instance, the auditory and the visual methods of teaching spelling, the teaching factor may be isolated and measured with a fair degree of accuracy in the following manner: Take two classes from the same school, or from different schools, which apparently have the same spelling ability as measured by some spelling scale such as the Ayres Spelling Scale. Combine the study period with the recitation period and have one class study spelling by the visual method, where the teacher

writes the words on the blackboard or has some other way of presenting them visually. In the other class the same words are to be studied, but all work is to be done orally and the children are not to see the words either in script or print for purposes of study. No work is to be done on the spelling lessons outside the classroom. At the end of one month a test may be given and the merits of the two methods determined. These conditions would be about as nearly controlled conditions as one could get in ordinary school-room procedure. Even here the difference in native capacities of the two classes, the difference in the two teachers, if two were employed, the difference in application of the pupils and many other factors would condition the results to some extent.

Our Educational Scales Have Been Subjective.—The problem for scientific education is to make our educational scales objective and universal. They must be so constructed that there can be no misunderstanding about them when they are placed in the hands of competent teachers. Educational scales make it possible to define educational products far more precisely than without them. Just as in the physical sciences it is far more nearly exact to say the temperature is 106 degrees Fahrenheit than it is to say, "It is the hottest day I ever saw," or "The day is boiling hot," or "It is an awful hot day," so in education we are refining our cruder terms and making them more definite. When an individual says it is the hottest day he ever saw, no one but the speaker knows what is meant, but when he says it is 106 degrees Fahrenheit any intelligent individual in any part of the world understands exactly what is meant. The expression, "the hottest day I ever saw," is as definite, however, as the 85 per cent the teacher gives a boy in penmanship, because no one knows but that teacher what it means. Before Thorndike made his hand-

writing scale educators were in the same condition with respect to handwriting as scientists were in respect to temperature before the discovery of the thermometer. In that day it was not possible to measure ordinary temperature beyond the cold, cool, warm, hot, and very hot, of subjective opinion.

We can easily measure the salaries of teachers because we have a scale of money price. We can measure the amount of time given by teachers because we have the best scale that the world knows, the scale of time divided into seconds, minutes, hours, etc. The most abstract thing in the world is a scale for length, weight, time, or units of temperature. As little as we care about scales they are among the most important things in the world. The skill and courage of the most daring seamen that ever traveled the seas, millions of them put together, do not do as much for the practice of navigation as the mariner's compass which is simply a scale for telling direction. While tests limit the "spread-out-ness" of our units of measurements, they do not give us a definite point on the scale. They do, however, give us narrow and definite boundary lines within which the measures lie.

Tests Do Not Indicate the Cause of Conditions.—A school achievement test, or any other kind, does not indicate what brought about the conditions found. It simply says that a certain state of affairs exists. When a physician's thermometer shows a temperature of 103° it does not indicate, in the least, whether the high temperature is caused by typhoid fever, influenza, diphtheria, or what not. It simply says that the patient's temperature is 103° . This is a result of perhaps many causes, and further diagnosis is necessary to determine what brought about this condition. In exactly the same way a score made in any subject is the resultant of many causes, of which teaching may be one. Teachers must use caution in

assigning conditions to definite causes unless it is absolutely known that the condition was brought about by the cause assigned.

How Standard Tests Differ from Ordinary Examinations.—Perhaps one of the best ways to show the difference between a standard test and an ordinary examination is to note, in a general way, the steps in making a standard test. To the uninitiated, school achievement tests look so much like ordinary examinations that teachers often wonder what the advantages of the standard tests are over the examinations.

The superiority of the standard test over the ordinary examination may be shown best, perhaps, by indicating the problems in making a standard test. The procedure is somewhat as follows: The first thing to be determined is the kind of a test desired; that is, shall the test cover several phases of the subject, or one? Shall it be a rate test, telling how much a pupil can do in a given time, or shall it be a difficulty test answering the question "How hard a problem can a child solve?" Or shall it be a quality test answering the question "How well can a child do a given task?" When the type of test has been chosen, then those making the test must decide upon the following points: Shall the test cover several parts of a subject as, for instance, the four fundamentals, common and decimal fractions, and percentage in arithmetic, or shall it be diagnostic and cover quite completely just one phase of the subject, as addition or division? Or, if the subject is language, shall the test cover in a general way all parts of speech, or shall it cover just verbs or pronouns? This having been decided the next point to determine is the principles on which the test questions shall be based. For instance, if a pronoun test is being made, shall the choosing of the test questions bear some definite relation to the type of errors people

make in the use of the pronoun forms, or shall something else be taken as a basis for choosing the questions? If more errors are made in the use of some pronoun forms than in others, shall there be more test questions on these particular forms, assuming that the test questions have been selected on some such basis as the above?

The next questions to decide are how many questions or parts shall go in the test and approximately how much time should it take to do the test. The number of questions or divisions is determined in part by the subject matter itself, and in part by arbitrarily fixing the length of the test. The approximate length of the test having been determined, the next thing is the actual drafting of the questions or parts and the submitting of these parts to a representative group of children usually from 2,000 to 5,000 in number. This is called the preliminary test upon which the standardized test may eventually be based. It may be, however, that the preliminary test will show that the mode of attack is not good and the whole thing must be "scrapped" and a new mode of attack adopted. Assuming that the mode of attack was satisfactory and that the preliminary test indicates that the desired information may be obtained by the methods adopted, the one making the preliminary test eliminates unsuitable material, modifies some of the items in the light of his experience, and brings out the test. He usually obtains some tentative standards; but the limitations of his time and money usually prevent his carrying this phase of the work to a satisfactory conclusion. There is, therefore, no test which does not require, after its publication, a thorough trial in order to set up norms of performance.

For many years workers in the field of educational measurements have been devising tests until now we have several for most of the subjects in the elementary school.

Many of the older tests have passed through their preliminary stages and may be said to be satisfactorily standardized. In some of the tests little attention has been given to questions of validity and the reliability of the measures which the tests yield. The availability of a number of tests for each of the common school subjects makes urgent a scientific determination of the validity and reliability of the respective tests in order that one may have information on which to base a rational selection.

In contrast with this long and tedious process, the traditional classroom examination consists of tasks of varying units of difficulty, worked at for varying amounts of time, and producing results of varying degrees of quality. The examinations do not measure any one thing. They measure a conglomerate of achievements conditioned by the three factors of quality of product, difficulty of task, and time consumed. Of course, standardized tests likewise measure a complex, but the methods of their derivation, the controlled conditions under which they are given, and the norms established make them far more reliable than the ordinary classroom examination.

Standardized tests are devised according to scientific procedure. The formulation of the questions or statements of the ordinary examination are based on the judgment of only one teacher and it often happens that the teacher does not give them very careful consideration. For the ordinary examination there are no standards. With the standardized test we have a statement of what scores the pupils of the several grades should make. The great problem in scientific education is to construct objective and universal scales, about the use of which there can be no misunderstanding when they are placed in the hands of competent teachers. One of the main

improvements due to the use of a definite set of standards is the elimination of the errors of prejudice.

It is to remedy these shortcomings of our marking systems that the scientific movement in education has devised tests and scales. Physical measurements require the thing to be measured to possess the quality of consistency and to remain constant while it is being measured. This is a fundamental necessity for logical thinking about measuring, counting, or enumerating. The things counted must all be of this same category. The thing measured must be constant in its character or composition so that one unit of it will be equal to any other unit of it.

The process by which the essential characteristic of consistency is obtained in educational measurements is the one used in physical measurements. It consists of distinguishing the possible controlling varying factors, devising means of holding all of them constant save one, and measuring that one. *This is the law of the single variable.*

A standard test generally has been given to pupils of many schools, which makes it possible to compare the scores made in one school with those made in another. Comparisons, however, must be made with care since many factors other than the teaching factor enter in.

Illustrating the Law of the Single Variable.—The operation of the law of the single variable may be illustrated in the methods used by Burgess in designing a reading test.⁴ The aim of this test was to find out how much printed material, of a *given level of difficulty*, a child could read “well enough for all practical purposes.” The attempt was to devise a test in which the child could readily succeed if he read well enough to grasp the impor-

⁴ May Ayres Burgess, *The Measurement of Silent Reading* (Department of Education, Russell Sage Foundation, 1921).

tant thought in each section, and in which he could not succeed at all unless he did comprehend each important thought. This was the interpretation which was put upon the phrase reading "good enough for all practical purposes."

The first step was to determine what the conditioning factors were in reading and choose one of these factors for measurement. The following is a list of 25 factors which the author of the test dealt with in measuring silent reading, with the disposition she made of them:⁵

To be measured:

Amount child can do in given time

To be eliminated:

Complex thought

Abstract thought

Technical thought and language

Catches

Puzzles

Accidental leads

Demands for spatial imagination

Irrelevant dramatic appeal

Ability to reproduce

Ability to remember

Ability to reason, or infer

Involved style

To be held constant through the test:

Memory span requirements

Attention span, multiple

Difficulty of action demanded

Time required for complying with instructions

Vocabulary difficulty

Sentence structure

Word arrangement

Amount of material to be read

Uniformity of print

Uniformity of space relations between pictures and print

Case of finding place on paper

Interest and corresponding effort on part of child

⁵ *Ibid.*, pp. 37-38.

Of course she could not entirely eliminate all the factors listed under the heading "To be eliminated," neither could she hold entirely constant the last group of factors in the above list. Of this sort are the factors: Amount of material to be read; difficulty of action demanded; and others which could not possibly be kept entirely constant. The purity of the final score, which was "*the amount a child could do in a given time*," was determined by the degree to which she was able to control some of these factors and eliminate others. Unless this could be done to a marked extent, her final scores would be a *conglomerate*, or a measure of many things instead of one, which measure would, in fact, be a measure of a complex.

As was said above, the essential characteristic of consistency may be obtained in educational measurements only by distinguishing the possible controlling varying factors, devising means for holding all of them constant save one, and measuring that one. This is the law of the single variable.

The reason why marks received by pupils in arithmetic and other subjects usually do not actually record their abilities and the value of their achievements is that they do not measure different amounts of the same thing.

The law of the single variable is a principle of measurement taught in the earliest school years and increasingly recognized in the comparative judgments of every-day life.

The tests designed to measure reading ability usually measure in addition other and different abilities.

Time of Day When a Test Should Be Given.—There is undoubtedly a best time in the day to give a school achievement test. If a high score is the thing desired, other things being equal, the morning, when the children are fresh, is better than the afternoon. Fewer errors are liable to be made in the early part of the school day.

There is another problem, however, to be taken into

consideration. It is the question as to how near we want to make the school conditions approach those in life outside the school. In our regular work-a-day life we cannot choose just when we shall do the difficult task. For instance, a railway passenger agent may have been working hard all day, and five minutes before six o'clock in the afternoon twenty-five people may rush up to the ticket window and demand tickets to some distant points. It may be that the train is already on the track ready to depart in three minutes. Those desiring tickets are excited because they fear the train will depart before they secure their transportation. Under conditions like these the ticket agent must answer their many questions, give them the proper tickets, answer the telephone, make change correctly, and attend to his other duties as ticket agent and telegraph operator. The fact that he has worked hard all day and is pretty well "fagged out" before the rush comes will not excuse him for any errors he may make as to the information he gives, the change he makes, or the tickets he sells. He is held strictly accountable for every act. It would have been to his advantage to have had the rush in the early part of the day when he was fresh and better able to cope with the situation. But he was compelled to accept it just as it happened to come. One criticism of our schools is that they are so different from extra-school life that children prepared in them are not able to cope with life's situations.

It is true that the children may be nervous and more apt to "go to pieces" in the latter part of the school day than in the forenoon, but that may be one of the reasons for giving the test to see how well they can perform under conditions which closely approximate those of extra-school life.

The child who hasn't been trained to withstand conditions of this kind would probably "go to pieces" when

he reaches adult life and attempts to fill a responsible position.

It is not maintained that one should pick out that part of the school day when the children are pretty well fagged out and give the test at that time, but, if that part of the day happens to be the most convenient time to give the test, it may not be wise to postpone it simply because the children are not as fresh as they are at some other time.

Number of Times a Test Should Be Given.—A test is given to a class for the first time to find out the condition of the class in reference to any subject. It indicates that the class has reached a certain state of proficiency in reference to that subject. It gives the teacher a point of departure. It shows the strong and weak points of the class. It indicates where work should be done. It too often happens, however, that teachers and superintendents give a test to find out the condition of a class in reference to a certain subject and then never act on the information gained. Such procedure is like a physician pronouncing a case typhoid fever and then making no effort to treat the case. The test should indicate what needs to be done. Teachers should act on the information given and test again to see what improvement has been made. There are some tests which cannot be repeated at short intervals because the children would remember enough from the first test to vitiate the results if it were given a second time with a short time interval. It is usually easy to avoid this difficulty, however, because two or more tests of equal difficulty are usually made in reference to any one subject. The practice effect is thus avoided. There are certain tests, however, such as the Curtis Standard Research Tests in Arithmetic, Series B, which may be repeated at short intervals without the practice effect vitiating the results. It would be a rare case for a pupil to remember the answers to any of these problems more than a day or two. Owing

to the nature of the Monroe Silent Reading Tests, however, the children might remember the correct answers to them for several months.

We give tests for the same reason that a merchant invoices his goods. He cannot determine his profits until he makes at least two invoices. Neither can a teacher determine the progress of her pupils until she makes at least two tests, the first one to determine a point of departure, and the second, or succeeding ones, to determine the progress made. Without the tests the teacher is simply guessing at the progress made, but she never really knows to what degree her efforts have been rewarded.

How Standard Tests Are Helpful in Improving Instruction.—Standard tests render assistance to the teacher in three ways. (1) Since the test has been constructed after a careful analysis and survey of the field, it gives a teacher a list of things that the pupil should be able to do. This is well illustrated by the Ayres Spelling Scale, which contains the 1,000 most frequently used words, and Monroe's diagnostic tests in arithmetic, which give a list of the significant types of examples in certain fields. (2) Since the tests are *standardized*, the teacher may know just what scores pupils ought to make. She is, therefore, given a definite objective aim to strive for in her teaching, an aim which the pupil can understand. The advantages of a definite standard are obvious. (3) Tests furnish the teacher with information concerning the abilities of her pupils. They point out phases of strength and weakness. With this information at hand she can plan instruction which will be more efficient since it will meet specific needs.

A test cannot be used properly unless it is accompanied by a complete set of instructions for giving it, for scoring the test papers, and tabulating the scores. For tabulating the scores a special class record sheet is usually provided.

The value of standard tests is realized through the use

that is made of the scores. They must be interpreted in *terms of the needs of the pupils for instruction*. This means doing more than determining whether the scores are above standard, at standard, or below standard. It means doing much the same sort of thing the physician does when, after he has ascertained the patient's pulse rate, temperature, and other symptoms, he prescribes treatment.

The use of standard tests should result in the improvement of instruction, and this is accomplished by means of the information which the tests yield concerning the abilities of the pupils. Growing children do not develop skill by instruction or personal exertion on the teacher's part. They develop best when they are inspired to voluntary effort. Mere repetition does not develop skill; it is repetition accompanied by a conscious desire to improve that brings results. Credit should be given for growth, not for the number of lessons completed.

It is easy to get a child to try once, but he will not keep on trying unless his efforts bring success. Standard tests automatically set for each child a task within his reach. He knows when he gets it done.

Speed tests are timed in order to "speed up" the children's work. Teaching a child to concentrate and to work efficiently does not mean prodding him to hurry. *Speed is to be acquired by study and practice*, not by special effort. The amount of work done in a given time is merely a symptom which may indicate to the teacher whether or not the child has studied the lesson sufficiently.

It is only by measuring the initial ability of children in the fall and the final ability in the spring that a teacher may know the progress made.

What Kind of School Achievement Tests Is Most Important?—Much discussion has been carried on as to the relative importance of various types of tests. It is

sometimes argued that the Courtis Standard Research Tests in Arithmetic, Series B, for instance, are not as valuable as other tests in arithmetic dealing with the more complex processes. To debate whether the fundamentals in arithmetic are more, or less, important than the more complex processes is analogous to debating which is the more important, the foundation of a house or the superstructure upon it. One cannot exist without the other. The four fundamentals in arithmetic constitute the foundation upon which other more complex processes are built and are absolutely necessary to the computation of these complex problems. The point is that we need tests in both fields. Tests in the fundamentals will not give information relative to the higher and more complex processes, and tests in the complex processes give information only indirectly as to the fundamental processes.

CHAPTER VI

THE CLASSIFICATION OF SCHOOL ACHIEVEMENT TESTS AND THE FUNDAMENTAL PRINCIPLES FOR DESIGNING THEM

School achievement tests show many lines of cleavage. Just as it is possible to classify school subjects as sciences, arts, and volitions, or as formal subjects, content subjects, and expression subjects, so it is possible to classify tests and measurements into many groups depending upon the particular features one has in mind when the classification is made. In order to bring out clearly some of the more basic principles in designing tests we shall note some of the classifications.

Diagnostic vs. General Tests.—One of the first problems that confront those making a test is whether or not the test shall be *diagnostic*. By a *diagnostic test* we mean one which furnishes a separate measure for each specific ability in the field of the test, or at least, of the important abilities. Scientific investigation has shown that in a subject such as arithmetic there is not simply one ability, but a large number of abilities. A pupil may be good in addition, poor in subtraction, and weak in multiplication. It is a rare case when a pupil is equally proficient in the several abilities. Each school subject, therefore, includes a number of abilities which are specific or distinct from each other to a considerable degree. The degree to which a test is diagnostic depends chiefly upon the amount of subject matter in that particular field. The Charters Language Tests for Pronouns, for instance, are diagnostic in that the

number of pronouns is limited and it is possible to give questions and examples using each of the pronouns in almost every conceivable way in which they are used in oral and written composition. Such a test may give a complete diagnosis of the child's language abilities in the use of pronouns and any shortcomings may be quickly located and corrected.

A *general test* is, in a sense, opposite to a diagnostic test. It yields average or composite measures of a child's ability in the subject matter in question. It may be illustrated by the miscellaneous language test devised by Charters or by the Courtis supervision tests in arithmetic. The latter is a single test covering the entire field of all the operations with integers. The test gives the pupil's general or average standing.

Grade norms cannot be used to make individual diagnosis. But we can see by them which children are below and which are above the level that they should attain in their grade. Grade norms will not give administrators what they most need to know, namely, which children have progressed at the rate normal for their age and native capacity, and which are performing at their maximum.

Degree to Which Tests Are Diagnostic.—The diagnostic characteristics of tests range all the way from zero, where the test is so general that specific abilities are completely lost sight of, to tests which may be said to be 100 per cent diagnostic. We may illustrate this fact by examining some of the handwriting scales. Both the Thorndike and the Ayres handwriting scales may be considered almost zero as far as measuring specific abilities is concerned. When these scales are used in measuring handwriting the sample to be measured is moved along the scale until the quality is reached that most resembles the sample to be measured. The specific characteristics of the sample are not taken into account. When the score is made up, it is a record of the

two of the abilities. This method was abandoned because it would require a choice among those parts for which the correlation is high and then among those where the correlation is not so high. This would resolve itself into an arbitrary choice which was the very thing he was attempting to avoid.

A second plan which suggested itself was to submit the list to a representative body of teachers and allow them to choose what phases should be measured. This plan was objected to on the ground that it was not known in advance how many elements must be chosen and it was doubtful if mutually exclusive points could be obtained in this way.

The plan finally adopted for determining what characteristics should be used in making a score card for writing was one which grew out of experience with the card itself together with the arbitrary choice of the author. A number of students were asked to grade samples of handwriting each week with the crude score card with the idea of determining what points should be used in order to give a complete account of writing. The students were cautioned to watch for points in the writing which were not covered by the card that they were using. In the light of the experience thus gained it was found that writing might be rather completely described under nine headings. These are: (1) spacing of letters; (2) spacing of lines; (3) spacing of words; (4) slant; (5) size; (6) alignment; (7) neatness; (8) heaviness; and (9) the formation of letters. Neatness, which is, in part, a function of other elements, was finally included to take account of such points as blotches, carelessness, and retracing.

Many of these headings were subdivided in order to increase the diagnostic value of the scale. The formation of the letters, for instance, was subdivided into: (a) parts omitted; (b) letters not closed; (c) parts added; (d) smoothness; (e) general form.

A diagnostic test of a different nature has been designed by Dr. Judd and was first used in the Cleveland survey. This is a series of 15 arithmetic tests, each composed of different types of examples. This series of tests has been called *spiral* because the abilities of the pupils are measured on successive levels of difficulty. The series is diagnostic rather than general. It yields measures of rate and accuracy with which the pupils do certain types of examples. The advantage claimed for the spiral method is that it offers a means for distinguishing errors due to accident from errors due to ignorance or incapacity. Errors of the latter sort will recur at regular intervals and may be readily recognized. Another argument for the spiral tests is the fact that time will not permit as many problems in a particular field as is desired so that each distinct mental operation may be thoroughly tested; hence, the device of a spiral arrangement, by which several related operations are combined in the same test in cyclic order.

The Courtis Standard Research Tests, Series B, are general tests in the field of each operation but diagnostic to the extent that they give information for each of the four fundamental operations; addition, subtraction, multiplication, and division.

Formal Tests and Reasoning Tests.—Tests may be classified, from the standpoint of the kinds of mental processes involved, as formal tests and reasoning tests. For measuring skill or automatic processes we use the former type. These tests measure immediate specific and preparatory outcomes of school training. The Courtis Standard Research Tests, Series B, are illustrations of formal tests. On the other hand, we may design tests to measure generalized outcomes of school training. The Stone Reasoning Test in arithmetic may be classed as an example of this kind.

two of the abilities. This method was abandoned because it would require a choice among those parts for which the correlation is high and then among those where the correlation is not so high. This would resolve itself into an arbitrary choice which was the very thing he was attempting to avoid.

A second plan which suggested itself was to submit the list to a representative body of teachers and allow them to choose what phases should be measured. This plan was objected to on the ground that it was not known in advance how many elements must be chosen and it was doubtful if mutually exclusive points could be obtained in this way.

The plan finally adopted for determining what characteristics should be used in making a score card for writing was one which grew out of experience with the card itself together with the arbitrary choice of the author. A number of students were asked to grade samples of handwriting each week with the crude score card with the idea of determining what points should be used in order to give a complete account of writing. The students were cautioned to watch for points in the writing which were not covered by the card that they were using. In the light of the experience thus gained it was found that writing might be rather completely described under nine headings. These are: (1) spacing of letters; (2) spacing of lines; (3) spacing of words; (4) slant; (5) size; (6) alignment; (7) neatness; (8) heaviness; and (9) the formation of letters. Neatness, which is, in part, a function of other elements, was finally included to take account of such points as blotches, carelessness, and retracing.

Many of these headings were subdivided in order to increase the diagnostic value of the scale. The formation of the letters, for instance, was subdivided into: (*a*) parts omitted; (*b*) letters not closed; (*c*) parts added; (*d*) smoothness; (*e*) general form.

A diagnostic test of a different nature has been designed by Dr. Judd and was first used in the Cleveland survey. This is a series of 15 arithmetic tests, each composed of different types of examples. This series of tests has been called *spiral* because the abilities of the pupils are measured on successive levels of difficulty. The series is diagnostic rather than general. It yields measures of rate and accuracy with which the pupils do certain types of examples. The advantage claimed for the spiral method is that it offers a means for distinguishing errors due to accident from errors due to ignorance or incapacity. Errors of the latter sort will recur at regular intervals and may be readily recognized. Another argument for the spiral tests is the fact that time will not permit as many problems in a particular field as is desired so that each distinct mental operation may be thoroughly tested; hence, the device of a spiral arrangement, by which several related operations are combined in the same test in cyclic order.

The Courtis Standard Research Tests, Series B, are general tests in the field of each operation but diagnostic to the extent that they give information for each of the four fundamental operations; addition, subtraction, multiplication, and division.

Formal Tests and Reasoning Tests.—Tests may be classified, from the standpoint of the kinds of mental processes involved, as formal tests and reasoning tests. For measuring skill or automatic processes we use the former type. These tests measure immediate specific and preparatory outcomes of school training. The Courtis Standard Research Tests, Series B, are illustrations of formal tests. On the other hand, we may design tests to measure generalized outcomes of school training. The Stone Reasoning Test in arithmetic may be classed as an example of this kind.

Rate Tests and Development Tests.—Another line of cleavage allows us to divide the tests into rate tests and development tests. Curtis and Rugg² especially use this terminology. Rugg makes three distinctions between these two kinds of tests. (1) Rate tests are distinguished from the development tests in that the latter make use of the "time" factor only incidentally or not at all. That is, the students are given practically all the time they need to do the test and their scores are reckoned only slightly, if at all, in terms of time. (2) The rate test differs from the development test in that the latter is made up of all kinds of subject-matter ranging from the purely formal and automatic material on the one hand, to complicated reasoning problems on the other. By rate tests we have in mind those types of tests in which "the ability involved in the working of any one problem is roughly the same as that involved in the working of any other."³ Of course, the ability involved in the solution of the various rate problems is not exactly the same as that involved in the solution of others in the test; but the same general type of mental processes is involved. (3) A third distinction is based on the organization or arrangement of the parts of the tests. In the rate tests one of two plans is followed. Either problems involving the same mental processes are grouped together in one test, or there is combined in one test a group of problems involving very closely related abilities. "The difficulty of each example in the test has been determined and the examples have been arranged in terms either of a rotating or a *cycle* principle, or all problems of the same difficulty have been put together."⁴ In the development test the difficulty of examples has been determined as in the case of the rate test, and the examples are

² *Scientific Method in the Reconstruction of Ninth-Grade Mathematics*, Supplementary Educational Monographs, Vol II, No. 1, 1918.

³ *Ibid.*, p. 65.

⁴ *Ibid.*, p. 66.

arranged in order of increasing difficulty. *Different kinds of abilities are measured* in the development test.

Quality, Difficulty, and Time or Amount Tests.—In general it may be said that the standard educational measurements fall into three clearly defined groups according to which of the three variables we seek to measure. They are tests and scales for quality of product, for difficulty reached, and for amount done.

1. *Quality tests* attempt to answer the question, *How well can a pupil perform a certain task?* Tests in composition, drawing, and handwriting are primarily quality tests. They are not scored on a basis of *right and wrong*; but there are all degrees of merit from the lowest to what might be called perfection. In writing, for instance, there is no *right or wrong*. The merit simply ranges from less good to more good through a continuous series of degrees of quality.

Reading is a classroom activity which does not readily lend itself to measurement by means of scales of quality. One reason for this is that it does not result in a tangible objective product which can be scrutinized and measured. Another reason is that quality in reading is an elusive thing which varies not only with different people but with the same person from moment to moment as he reads.

2. *Difficulty tests* attempt to answer the question, *How hard a task can a child do?* They are measured in terms of *right and wrong*. The question may be, How hard a word can a child spell? How difficult a problem can a child solve in arithmetic? In scales for difficulty, the variable which is measured is the difficulty of the work which the child can do. The difficulty of successive tasks is carefully increased and controlled and the child is allowed to overcome it if he can. The quality of his work must be high enough to be considered "right."

The commonest of all classroom questions is probably that

which relates to the difficulty of the work which the child can do correctly. The usual method which has been adopted to answer these questions is to prepare a series of tasks carefully graded in difficulty. Those near the beginning of the series are so easy that almost any child in the group can do them. As the series progresses the questions become increasingly more difficult. In scales for difficulty the amount of time allowed for the test should have no effect on the score. The independence of time must hold, not only for most difficult problems near the end of the series, but also for easy problems. There are some types of difficulty problems that a child can answer correctly at once or not at all. Spelling illustrates the point. If a child cannot spell a word immediately, no amount of time will aid him in his efforts. Spelling and arithmetic, and, less definitely, geography, history, and grammar constitute appropriate subject-matter for difficulty tests. Some of these are informational subjects in which, by the common verdict of society, the information is only valuable if it is accurate and correct. A type of handwriting that is somewhat inferior to another sample may be of almost equal practical value. The same cannot be said of spelling, arithmetic, history, geography, or grammar. Classroom products in these subjects are judged as *right* or *wrong*.

The student who devises a scale for difficulty, then, must either present evidence to show that scores for the particular ability he seeks to measure are not affected by differences in time, or he must devise methods by which the amount of time the pupil is allowed to spend on each task within the series may be controlled and recorded. The development tests discussed above have many characteristics of the difficulty tests here under consideration, but are not like them in every particular.

3. *Time tests, or tests for the amount done*, are, in reality,

the rate tests discussed above. Some additional characteristics, however, will be given under this heading. It is to be noted that *time and amount are complementary terms*, each of which depends upon the other for its meaning. Time implies amount, and amount implies time. The question, How much can be done? demands a statement of the time allowed for doing it, and the question, How long will it take? depends upon how much there is to do. This variable may be handled in two ways: (1) The problem may be to determine how much can be done in a given time as in the Courtis Arithmetic Tests, Series B. In this test the time is eight minutes for addition, for instance, and the information sought is, How many problems can be solved in that time? (2) A definite amount of work may be given and the problem is to determine how long it takes to do it. The latter method would not work well where the test is given to a large group of children at the same time, since their finishing times would be different and therefore hard to record. Quality, difficulty, and time or amount tests may be illustrated by the athletic contests we carry on in our schools. Marksmanship is a measure of quality answering the question, How *well* can one shoot? The other variables of difficulty and time are kept constant. The high jump is a measure of difficulty; quality, which is *good enough to clear the bar*, and time are kept constant. The 100-yard dash is a time test.

The three questions, *how well*, *how hard*, and *how fast*, represent the teacher's attempts to measure the three fundamental factors of *quality*, *difficulty*, and *time or amount*. The educational tests and scales that have been devised during the past ten years are attempts to help her answer these questions and each of them seeks to measure some one of those same three fundamental factors.

Measurements by Opinion, Comparison, and Standardized Tests.—The methods for the measurement of school

achievements may again be classified according to other principles as follows: One method, and the one generally in vogue, is that of personal opinion. Measurements by this method are valuable just to the extent to which the persons passing judgment are qualified to give expert opinion. As was noted in the introductory chapter, personal opinion, even though expert, is worthless in the face of facts unless the opinion happens to agree with the facts.

A second method of measuring school achievements is by comparison. We may say that one pupil, or class, or school, ranks fifteenth as compared with 25 other pupils, or classes, or schools. This method has merit, but the chief criticism of it is that it does not measure in reference to a standard. Even though a class, or school, ranked first among 25, it might still be a poor class or school. The ranking or comparison method is used to a greater extent than most teachers are aware of, however, when monthly report cards are made up. They usually reason that *A* is a better student than *B*, therefore *B* is given a grade of 80 per cent while *A* is given 85 per cent.

Measurement by comparison is based on the fundamental idea that the common practice is the result of the judgment of many men who have attempted to solve the same, or very similar problems. The order of merit method is used where units are not definite.

A third method of measurement and the one about which we are primarily concerned here is measurements by scientifically determined standards and units. This is the greatest contribution which has been made to education in the last twenty years.

Classification by Educational Tests.—Thus far we have discussed the general classification of tests. We shall now note still a different group of educational measures and

how they may be used in the reclassification of pupils in the schools, and also how they may be used to measure educational processes and products. It is obvious that if we continue to teach pupils in classes we must so classify them that those of approximately the same mental ability will be together. Experiments made seem to indicate that approximately 25 per cent of the pupils in any grade belonged mentally to a lower grade and about 25 per cent belonged to a higher grade. If the teacher does justice to the upper 25 per cent, he is over-teaching the other 75 per cent, or if he addresses the average pupils, he is under-teaching the upper 25 per cent and over-teaching the lower 25 per cent. Fortunately, we are developing methods for the reclassification of pupils which will make the pupils in the various classes more nearly homogeneous. Both intelligence and school achievement tests are employed in this reclassification. Franzen, McCall, and others are making use of the educational quotient (E. Q., educational age divided by chronological age) as a means of reclassifying students and measuring their school progress.

In order to compute a child's educational age in any subject, it is necessary to give a series of tests to a large number of pupils and determine the norms for children of different ages in that subject. For example, suppose an additional test were given to a large number of children ranging in age from eight to fourteen years. Suppose further that the average number of problems solved by children chronologically eight years old was four; those eight years, six months old, six; those nine years old, eight; those nine years, six months old, ten, and so on. Then a child doing six problems would be considered eight years, six months old educationally in that test, irrespective of his chronological age. If a great number of tests were given to a child in the subject of arithmetic

and a composite score were computed and compared with the norms for children of the various ages, a child's educational age in arithmetic might thus be determined. In a similar way a child's educational age may be computed in other subjects.

Educational Age and Mental Age Compared.—It is clear that educational age would be a better measure for the classification of pupils than mental age because the former represents the actual accomplishment of the pupil—the pupil's educational status—while the latter simply shows potential ability. The factors determining educational age are both hereditary and environmental. At the present time we have fairly well-established grade norms in the various subjects but do not have educational age norms. Suppose educational age norms were established, how could we utilize them in the classification of pupils? In what ways are they superior to mental ages? If age norms are used, it will be possible to reclassify pupils and put each one on his proper educational level. When the educational age is used as a basis of promotion or demotion, it takes cognizance, not only of the pupil's mental age which is potential, but also what he has actually done, which is one of the best ways of prophesying of what he is able to do.

If educational age is to be used in the reclassification of pupils, how much better than the average of his class must a pupil be before he is allowed to skip a grade and enter a higher class? Or, how much poorer than the average of his class must he be before it is considered wise to fail to promote him or even to demote him? Here both the E. Q. and I. Q. are taken into consideration. As was indicated above, the best way to prophesy what rate of progress a child can make is to find what rate he has made. There is a rather high correlation between the educational age of a pupil and his mental age. Only

tentative answers can be given to these questions since not enough work along these lines has been done to be entirely sure as to what is the best thing to do. It is the opinion of some investigators that the reclassification of pupils in a relatively small school should be somewhat as follows: No pupil should be allowed to skip a grade unless his educational age exceeds the educational age of the lowest 25 per cent in the grade to which he proposes to enter. In the matter of demotion and failure to promote, no child shall be denied his normal promotion nor be demoted unless his educational age falls within the lowest 25 per cent of his class. These rules not only take cognizance of what the child has done but also his potential power. They do not mean, of course, that simply because a child belongs to the lowest 25 per cent of his class, he shall be demoted, or because he exceeds the average of his class, he shall skip a grade, but they mean, if a pupil were an average pupil in class, that he would not be denied promotion nor be demoted.

Just as a mechanic working in a garage has a rather definite procedure in diagnosing engine trouble, so an educational expert soon develops a rather definite procedure in diagnosing educational ills. When a child fails in promotion, one of the first things is to get his mental age. There is usually a substantial correlation between mental age and school marks, and a child's mental age may throw much light on his failure to be promoted.

Accomplishment and Educational Quotients.—The *accomplishment quotient* (A. Q.) of a pupil is found by dividing his educational age by his mental age. It shows the degree a pupil's actual progress approaches his potential progress. This is perhaps the best measure we have of the instruction and the application of the pupil. When a report of this kind is taken home, the parent may form some intelligent idea as to whether the pupil is working

up to his optimum capacity. He may find out whether the pupil is progressing at a rate normal for his mental and educational ages. This also becomes a protection to teachers. If they can show that all the pupils under their guidance have progressed at a rate normal to their general intelligence, they have a good defense for adverse criticism of their teaching.

Another unit of measure that is being employed in measuring educational processes and products is the *educational quotient* (E. Q.), which is the educational age divided by the chronological age. It is the division of what is, by what it would be if the pupil were normal. It gives the percentage of normality. The quotient thus derived will indicate whether the pupil has made normal progress for his age or whether he has progressed more rapidly or slowly than the average. Indirectly, at least, the educational quotient throws some light on the pupil's general intelligence. A high educational quotient usually signifies a high intelligence quotient, but not always. Well taught pupils with a low I. Q. may have relatively high educational quotients.

When pupils are classified according to their native capacities and educational ages, we may then begin to judge the quality of the teaching more intelligently. A reasonable increment or interest on the mental capital invested will be all that is expected of teachers. Educational norms will serve as goals for teachers and pupils. Parents and supervisors will not expect increments out of proportion to the mental capital invested. On the other hand, teachers will know when pupils are doing an honest day's work, that is, when they are accomplishing a normal amount for their general intelligence and educational level.

Principles for the Choice of Subject-Matter for Educational Tests and Scales.—What shall be the controlling factors which will determine the type of subject-matter

used in tests? Many factors enter into this problem, some of the more important of which we shall discuss here. As was discussed in the previous chapter, a scale is a ladder, or a linear rule, extending from the worst to the best, from the lowest to the highest, or from the easiest to the hardest, and indicates the steps or degrees by which intermediate achievements may be gauged. Now it is evident that one making a scale of progressive degrees of difficulty may seek subject-matter with only this idea in mind and pay no attention to the practical or utilitarian value of the subject-matter that enters into the scale. This method is sometimes called the *statistical* method as opposed to the *analytical* method, which does take cognizance of the subject-matter field from which the material is to be chosen. In Woody's arithmetic scales, for instance, he states that his "fundamental idea was to derive a series of scales that would indicate the type of problems (examples) and the difficulty of the problems (examples) that a class can solve correctly."⁵ This method assumes that an example is suitable for use in a test simply because it is done correctly by a gradually increasing per cent of pupils as one proceeds from grade to grade. He selected his test material not on the basis of its arithmetical significance but on the basis of the consistency of the pupils' reactions. It seems to the author that his procedure is open to criticism on the ground that arithmetic is a tool subject and there is little use in testing the ability of pupils to use a tool that they will rarely or never be called upon to use in life outside the school.

It seems that in the tool subjects, at least, one must determine the subject-matter for tests and scales very largely on the basis of its utilitarian value.

⁵ *Measurements of Some Achievements in Arithmetic*, Teachers College Contributions to Education, No. 80 (1916), p. 1.

Information Desired Determines the Type of Subject-Matter.—Tests are of three general types, difficulty tests, speed tests, and quality tests. The kind of subject-matter used in the test will be determined by the kind of information desired. If a teacher wants to know how rapidly a pupil can do arithmetic, for instance, the problem is one of choosing subject-matter of a constant level of difficulty and of uniform quality, which is that quality which the public has been accustomed to call *right*. The subject-matter should be chosen from fields that supply the tools for doing society's work in arithmetic. The rate test will differ from the difficulty test since in the former but one type of subject-matter is usually chosen, while the difficulty test usually includes several types.

SOME CHARACTERISTICS OF AN IDEAL EDUCATIONAL SCALE

An ideal educational scale must have at least the following characteristics: (1) it must have an accurately defined zero point; (2) the steps above the zero point must be of equal magnitude; (3) the scale must measure the desired educational product; (4) it must be so simple in its application that it is adapted to the classroom; (5) it must not require an undue amount of time in administration. We shall note briefly the significance of each of these characteristics.

1. The Establishment of a Zero Point.—In education as in the physical sciences two positions may be taken as the zero point on the scale. The first may be called the absolute zero, which means *just not any* of the thing in question. In measuring heat, for instance, absolute zero is 273 degrees below the point we ordinarily call zero on the centigrade scale and means that all the heat has passed out of the thing in question. Establishing an absolute zero point in education is a difficult thing to do. The zero

point on the Hillegas Composition Scale was determined by 40 professors of English, editorial writers, psychologists, and educational experts. The composition with zero merit reads as follows:

Dear Sir: I write to say it aint a square deal. Schools is I say they is I went to a school. Red gree green and brown aint it hit to a bit I say he don't know his business not today not yesterday and you know it and I want Jennie to get me out.

That is probably a little better than zero. There is a little merit concealed in it but not enough probably to prejudice the matter seriously. The fact that zero points do not stare us in the face in the case of mathematical originality, or knowledge of German, or ability in writing as they do in the case of measures of length, and weight, and time, is no excuse for not trying to get them. If we get scale points defined and their distances defined and established in reference to an absolute zero, there is no further difficulty in constructing a scale to measure mental achievements. Such scales have every logical qualification that any of the scales in the physical science have.

Zero points on scales are imperfectly known, and as a result we add and subtract educational quantities with much less precision than desirable. We cannot say that one product is twice as good as another, or one task twice as hard as another, or that one improvement is twice as great as another unless we establish a zero point. Statements of these kinds are intricate and subtle matters involving presuppositions which must be kept in mind.

The ordinary scale for weight exemplifies an ideal scale in four respects. First, it is a series of perfectly definable facts. All men the world over know exactly what is meant by two grams, four grams, etc. In the second place, each amount is a definite amount of the same kind of thing. In the third place, the difference between any two of the

amounts is perfectly defined in terms of some unit of difference. The step from four to five grams is the same as from six to seven. Lastly the zero point of the scale is absolute. That is, it means *just barely not any* of the thing in question.

Thorndike acquired an actual zero produced by a human being in penmanship. He has a signature of a letter which cannot be read *in toto* and in which no letter can be read by any one of hundreds who have tried. He defined zero as the suppositional handwriting such that, though recognizable as handwriting, it has no legibility, no beauty, no value as penmanship. When we used zero in education in the past, we usually have had in mind a relative zero. The value of this zero is usually a subjective one, hence, ill-adapted to measure educational products.

Education is having the same difficulty that the natural sciences pass through in the standardization of their units. In the matter of recording temperature, for instance, scientific progress was handicapped by the fact that different individuals were using different reference points when measuring temperature. Finally, after long and costly delays, the methods of measuring temperature was reduced to two competing systems, one of which took the freezing point of water as the point of reference and the other took a point 32 degrees below that point. In measuring the height of land forms scientists agreed to take sea level as the point of reference. Other things might have been taken but they would not have been so universally applicable, and, if a number of different things had been taken, much time and labor would have been needed to convert one unit into another in order that comparisons might be made.

In education the tendency has been to search for some absolute zero point for the trait being measured. This is obviously a difficult task and results in much confusion, even assuming that it can be scientifically determined, be-

cause each particular test will have its own zero point. Many methods are used in the location of these zero points and points of reference. McCall⁶ mentions six methods of locating the zero points in scales now extant: (1) the reference point on unscaled tests is *just no score* on the material of the particular test; (2) the zero point is guessed at by the author of the scale; (3) the reference point on judgment scales is the median judgment of judges as to the location of zero merit in composition, handwriting, and the like, as in the cases cited above; (4) the zero point is located by the use of the per cent of pupils in some early grade who make no score on very easy material; (5) the reference point for other scales is three times the standard deviation⁷ below the mean of the group for whom the test was devised; (6) the reference point is simply the lowest score made. There are still other methods for locating points of reference in scale building.

Since there is a lack of agreement as to just what the zero point in reading and other subjects should be, McCall proposes to take as the reference point for school achievement tests, not a zero point, but the mean performance of children between the ages 12 and 13 years. He thinks that such a point could be used for any mental trait regardless of the location of its absolute zero, if such there be.⁸ He would then measure the school achievements of children in other grades in terms of the 12-year-old children. The grades ranged from 5 S. D. (standard deviation) below the mean score of the 12-year-olds, the mathematical zero of the scale, to 5 S. D. above the mean score of the 12-year-olds. Each S. D. was divided into 10 units, making the

⁶ "Proposed Uniform Method of Scale Construction," *Teachers College Record*, Vol. 22, No. 1, Jan., 1921, pp. 31-51.

⁷ See Chapter X for a definition of standard deviation.

⁸ *Op. cit.*, p. 43.

entire range from 0 to 100 with 50 as the mean score of the 12-year-old children.

By this method any pupil who makes a score of 50 has an ability equal to the mean ability of the 12-year-old children, and a pupil who makes a score of 40 has an ability of 10 units, or one S. D. below the mean ability of 12-year-olds. A pupil with a score of 75 is 2.5 S. D. above the mean ability of 12-year-olds and so on.

McCall gives four reasons why the mathematical zero is located 5 S. D. below the mean instead of at the mean: (1) this procedure eliminates cumbersome plus and minus signs; (2) it forms a convenient range of points between 1 and 100 with the reference point at the easily remembered 50; (3) this procedure carries the scale down and up as far as any one will need to go; (4) it gives a mathematical zero which is close to the supposed absolute zeros for reading, spelling, writing, composition, completion, and other typical mental functions.

One objection made to this scale is that the "times statement" cannot be employed in dealing with mental traits because the absolute zero point is not found. That is, one cannot say that John has three times the ability of James in a certain subject unless the absolute zero point of ability is determined. Nevertheless, 5 S. D. below the reference point cited above gives a mathematical zero that corresponds reasonably well with the absolute zeros determined in most scale making. And, when all is considered, the best way to appreciate ability of an individual is to refer him to the mean ability of his own or some standard group.

Another objection urged against this point of reference is that any score above or below this point would not indicate whether the pupil possessed much or little of a particular trait, because a 12-year-old pupil might possess little of a certain trait and relatively much of another.

This defect is to be remedied by the use of the absolute zero of the trait, provided such can be found.

The time of birth is used in most scales for measuring general intelligence as the point of reference. That is, in attempting to measure the general intelligence of an individual, his score is measured in terms of years and months of mental age.

McCall would not only standardize the points of reference in mental measurements but he would also standardize the units, and in each case would use some function of the variability of 12-year-old children, preferably the standard deviation. Thorndike and his students have constantly used some function of variability as the unit of measure.

The Binet-Simon Scale revised by Terman has met the qualifications for scientific scale building in that it has a definite tangible point of reference, the time of birth; it is simple and objective and is easily understood. But it fails to meet one condition for an ideal scale in that the units of the scale are not equal. The development of general intelligence is measured in years and months. The interval between eight and nine years, for instance, is larger than the interval between 14 and 15. In certain traits the unit above the age of 16 becomes zero. Because of the effects of social conditions, it becomes difficult to build up a scale on the age basis which will measure satisfactorily the general intelligence of pupils below the age of eight and above the age of 12. Hence, a scale of this kind cannot satisfactorily score pupils with exceptionally low or with exceptionally great ability. Judgment scales discussed above may be converted into scales of this kind thereby making all scales performance scales.⁹

⁹ For a more elaborate discussion of the T scale the reader is referred to the "Proposed Uniform Method of Scale Construction," by McCall, *Teachers College Record*, Vol. 22, Jan., 1921.

2. Making the Steps of Equal Magnitude.—Three methods are in common use in making the steps of a scale of equal magnitude.

(a) *The Method by Competent Judges.*—Many things in education must of necessity be left to the consensus of opinion of competent judges. The relative merits of two drawings, for instance, must always be determined in this way. Or, if we do not use the functional method in handwriting, the relative merits of the various samples to be measured may be determined by competent judges. By this method we may say that differences in merit between samples of handwriting, for instance, are equal when they are noticed by an equal number of competent judges. For example, if we had 1,000 of the best judges in handwriting in the world, and 750 of them were to judge a certain sample of handwriting designated by the figure 10 as better than another sample designated as 9, and the same number would say that sample number 11 was better than sample number 10, then we might say that sample 10 is as much better than 9 as 11 is better than 10 because the differences were noted by an equal number of competent judges.

This method rests on the fundamental assertion that *equally-often-noted differences are equal*. Owing to the nature of education, it is probable that a large part of the pedagogical scales of the future will be based on the consensus of judgments by competent judges. In drawing, English composition, music, handwriting, and such subjects, it is practically impossible to measure the results save by means of scales thus designed.

It should be noted in passing that the whole theory of scale development may be classified under two general methods: the *judgment method*, as the Thorndike and Hillagas scales, and the *ratio method*, followed by Ayres in making a spelling scale, and by Trabue in making a language scale. It may readily be shown that the school

subjects are susceptible of arrangement in a certain serial order which will indicate the method of scale derivation applicable to them. At one end of the series will be such subjects as spelling and arithmetic, which lend themselves to the ratio method, that is, to the expression of the relation between the actual number of correct responses and the possible number of correct responses. At the other end of the series are such subjects as composition and penmanship. Scales for these subjects must be derived by the judgment method. One sample of handwriting is better than another not as a fact but as an opinion. A specimen of English writing is better than another precisely because competent judges think it is better. Between the two extremes of school subjects are a number of subjects such as history, geography, and literature, which may be measured by either method.

(b) *By the Functional Method.*—By this method the quality of the thing is not measured directly, but indirectly by measuring the degree to which the particular thing or product functions. The Gettysburg Handwriting Scale is an example of this kind. By this method two samples of handwriting were said to be equally legible if readers could read one sample as rapidly as the other. The method of making the steps equal is as follows: Suppose three samples of handwriting are being considered, of which expert readers can read sample *A* at the rate of 100 words per minute, sample *B* at the rate of 120 words per minute, and sample *C* at the rate of 140 words per minute; then we may say that sample *B* is as much better than sample *A* as sample *C* is better than sample *B*, and that therefore the steps are equal.

Measures in the physical sciences are quite often made not by measuring the thing directly but by measuring something that varies with it. For instance, we do not measure heat directly but measure the length of a mercury column

which we know varies directly with the amount of heat in the body. By this method we know that the amount of heat that it takes to raise a column of mercury from a point marked 10 on the scale, for instance, to one marked 11, is the same, approximately at least, as the amount of heat that it takes to raise the mercury column from 11 to 12 or from 15 to 16. By the same principle, Ayres established approximately equal steps on his writing scale by assuming that progressive degrees of merit varied directly with the "readability," or the speed with which the various samples might be read. To the physicist those differences are equal which are produced by the same cause under the same circumstances, or under which the same conditions produce the same effect.

For a long time we measured fatigue by measuring the distance between the points of the æsthesiometer when placed on the skin and, although the correspondence between cutaneous insensitivity and fatigue has been more or less discredited, it is not discredited on the ground that the fatigue element could not be measured in this way, provided there is a correspondence.

Perhaps it should be noted in passing that none of these scales approaches perfection. They are still crude, but much better than the old methods based on personal opinion.

The Ayres Gettysburg Handwriting Scale has been criticized especially on the ground that in making the scale the different qualities of handwriting were determined by the rapidity with which the samples might be read, but when the scale is used in the school room the merit of a particular sample is determined not by how rapidly it may be read but by the nearness to which it approaches in form and general appearance a sample the readability of which is known. That is, the scale was made from a functional standpoint, the speed with which the various samples might

FIGURE IV. MEASURING SCALE FOR ABILITY IN SPELLING
(After Ayres)

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
99	98	96	94	92	88	84	79	73	66	58	50	2nd	58	50 3rd	73	66	58	50 4th	50 5th	50 6th	50 7th	50			
3rd	100	100	99	98	96	94	92	88	84	79	73	66	58	73	66	58	50 4th	73	66	58	73	66	58	50 7th	50
				4th	100	99	98	96	100	7th	100	8th	100	99	98	96	94	92	88	79	73	73	66	58	50
						5th	100	6th	100																

be read; but when a sample is to be measured, its quality is determined not by the speed by which it may be read but by the nearness it approaches in form and general appearance to a sample on the scale.

(c) *The Proportion-of-Pupils-Solving Method.*—In a spelling test, for instance, the words may be ranked according to spelling difficulty by determining the number of children that fail to spell them. For example, if three of the words are home, church, and separate, and 95 per cent of the children are able to spell the word, home, 90 per cent are able to spell the word, church, and 85 per cent are able to spell the word, separate, we may be assured that the words are arranged according to their spelling difficulty from the easiest to the most difficult. It does not follow, however, that since 5 per cent fewer pupils were able to spell the word, church, than were able to spell the word, home, and since 5 per cent fewer were able to spell the word, separate, than the word, church, that the increase in spelling difficulty from home to church equals the increase from church to separate.

In order to make the steps of equal difficulty, the normal distribution curve is brought into use. We may illustrate how the steps are made equal, or approximately so, by reference to the headings of the Ayres Spelling Scale, Figure IV. Dr. Ayres divided the words in his spelling scale into 26 divisions or columns. The words in each column are of approximately equal spelling difficulty, and the steps in spelling difficulty from each column to the next are approximately equal. The figures at the top of the scale indicate the approximate average scores of correct spellings that may be expected among children of the different grades and of the same grade. Thus, in column K, for instance, the 58 at the head of the column is the average score that should be expected from second-grade pupils attempting to spell the words in this column. The average

score for third-grade pupils is 79, for fourth-grade pupils, 92, and so on. The numbers 99, 98, 96, 94, 92, etc., at the top in Figure IV for the second grade are as near the mid-points of the equal steps as can be obtained without using fractions. That is, the step from column *A* to column *B* equals the step from column *B* to column *C*, and so on.

In making the spelling scale he assumed that the spelling ability in any one grade is distributed according to the normal probability surface.

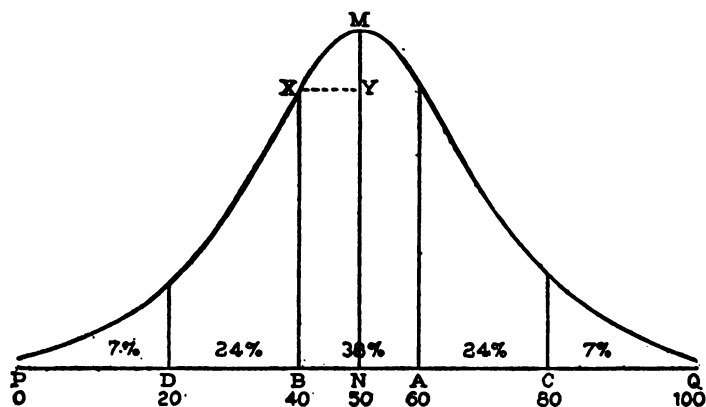


FIGURE V. ILLUSTRATING THE DISTRIBUTION OF SPELLING ABILITIES
(adapted from Ayres)

That is, taking any grade, as the third, for instance, and representing it by a normal distribution curve as in Figure V, we note that at the extreme left, the curve is very near the base line which indicates there are very few exceptionally poor spellers. In the middle, the curve is the greatest distance from the base line thus representing a large proportion of medium spellers. The median line, in Figure V, represents the 50 per cent in the third grade, Figure IV. The horizontal line, *xy*, from the median to the curve represents 1 sigma distance (sigma, σ , is the unit

for standard deviation) and intersects the curve at a point at which it changes from convex to concave.

This distance is always a constant function of the curve of normal distribution and in the Ayres study was chosen as the unit of measure along the base line. He laid off on the base line to the left of the line, *MN*, a distance equal to 2.5 sigma. The part from *N* to *B* is 0.5 sigma, from *B* to *D* and from *D* to *P* is 1 sigma each. Since the curve does not meet the base line at 2.5 sigma from *N*, but theoretically meets it only in infinity, we may assume that since only 0.62 per cent of the area between the curve and the base line lies to the left of point *P*, that for practical purposes it meets the base line at 2.5 sigma from *N* and this point may be considered as zero.

Ayres thus divided the base line into five equal parts. He called the extremes from left to right, 0 and 100 respectively. Assuming, as he did, that the entire frequency might be included between points 2.5 sigma to the left of *N* and 2.5 sigma to the right of *N*, he found that 7 per cent of the entire area between the curve and the base line lies to the left of the line erected at point *D*. Between this line and the perpendiculars to the base line at point *B* are 24 per cent of the cases. Between the lines erected at points *B* and *A* are 38 per cent of the cases; between *A* and *C* are 24 per cent of the cases; and between *C* and *Q* are 7 per cent of the cases. In scaling the words he found that 7 third-grade children out of 100 failed to spell the word "has," while 93 per cent spelled it correctly.

Applying this fact to the curve, the word "has" would be located at point 20 on the base line which would have 7 per cent of the cases to the left of it and 93 per cent at the right. In a similar way, a word missed by 31 per cent of the children would be located at point *B* on the base line. By the same method all the words were thus located in the spelling scale.

By dividing each of the five divisions on the base line into five equal parts, Dr. Ayres made a total of 25 steps ranging from 0, or near 0, to 100 in his spelling scale. The average of all the values that might theoretically be contained in each of these 25 steps has thus been determined to the nearest whole number and this value has been assigned to the step. These 25 values are 100, 99, 98, 96, 94, 92, 88, 84, 79, 73, 66, 58, 50, 42, 34, 27, 21, 16, 12, 8, 6, 4, 2, 1, 0. The limits of these values are as follows:

Score	50	means	from	46	to	54
"	58	"	"	55	"	62
"	66	"	"	63	"	69
"	73	"	"	70	"	76
"	79	"	"	77	"	81
"	84	"	"	82	"	86
"	88	"	"	87	"	90
"	92	"	"	91	"	93
"	94	"	"	94	"	95
"	96	"	"	96	"	97
"	98	"	"	98	"	98
"	99	"	"	99	"	99
"	100	"	"	100	"	100

Noting the scores 100, 99, 98, 96, 94, etc., in the top row and at the left in Figure IV, we see that the steps do not appear to be equal since they are 1, 1, 2, 2, 2, 4, etc. In order to make the steps of spelling difficulty equal from one column to another the average scores are computed in per cent and transmuted into units of standard deviation. This may be done by referring to tables for converting the per cent failing into units of standard deviation. We shall note more specifically the curve in Figure V to illustrate two measures of deviation that are used in making the steps equal. Theoretically, the curve reaches the base line only in infinity. The area between the curve and the base line is divided into 10,000 equal parts and each

particular section carefully mapped out so that it is known exactly where lines perpendicular to the base line must be erected on each side of the line MN to include the middle half of the area of the figure, or any other fraction of it. When lines are erected at equal distances from the line MN so that the area between the curve, the base line, and these two perpendicular lines is equal to one-half the total area, the distance of these perpendicular lines from the line MN is known as the *probable error*. In other words, the probable error is a distance on either side of the measure of central tendency (MN) that will include the middle half of the measures. It is evident that the line PQ may be divided into any desired units of probable error (or P. E., as the units are usually designated). It is also evident that a unit's distance on the line PQ near the central part of the figure would include a much larger area than the same linear unit would out near the ends of the curve. The area included between two perpendicular lines erected at units distance apart at some column as N , Figure IV, for instance, may be eight or ten times as great as the area included between two lines similarly drawn at or near columns Y or Z .

Lines erected at equal distances from the line MN so as to include the middle 68.26 per cent of the area between the base line and the curve are said to be erected at a distance of one sigma (σ) from the line MN . Sigma (σ) represents the standard deviation. It is evident, therefore, that the line PQ may be divided into any number of equal parts, each part being represented by sigma or a fraction thereof. As in the case of P. E., two lines erected perpendicular to the base line near the center of the curve and at sigma's distance apart would include a far greater per cent of the area than if erected at sigma's distance apart but near the end of the curve.

Now if one transmutes the per cents given at the head

of each column in the Ayres Spelling Scale for any particular grade, as for instance the third, into units of standard deviation, or sigma, it will be found that in passing from left to right the words grow progressively more difficult by approximately equal steps, and that the size of the steps is 0.2 sigma. Standard deviation is discussed more fully in Chapter X.

3. The Scale Must Measure the Desired Educational Product.—It is not always easy to tell just what a test measures after it is given. For instance, it is very doubtful just what the Trabue language tests measure. They may measure general intelligence or language ability or both. The scores are very difficult to interpret. Suppose a pupil makes a poor score on these tests. What shall the teacher or the administrator do as a result of it? Because of their general nature, little guidance comes to the teacher from tests of this kind. On the other hand, if a pupil makes a low score on the Charters pronoun tests, the teacher knows immediately where more drill work should be done. It gives her a point of departure. She knows that she has tested the pupil in a specific field and knows his weak points. When scores are a result of many variables it is difficult to tell just what the test has really measured. Care should be taken, therefore, to see that the test really measures what it is designed to measure.

4. The Test Must Be so Simple in Its Application that It Is Adapted to the Classroom.—The only way that it is possible to determine the intellectual achievement of a pupil is by what he does. In school achievement tests, pupils are asked to do a great many things, and it is evident that if the directions for giving the tests cannot be well understood by the pupil, the product will not be a correct measure of the pupil's ability. Any one who has given tests knows that even when the instructions are so simple that it would seem almost impossible for a pupil to mis-

understand them, yet, in a class of 40 pupils there will be one or two who do not know what the test calls for.

The ideal test must be of such a nature that the recording of answers, or the execution of the design, if a drawing is called for, may be done with the minimum amount of time and energy, so that practically all the effort will go towards the solution of the questions or problems rather than recording the answers when once worked out and made.

5. Tests Must Not Require an Undue Amount of Time in Administration.—Tests that require an undue amount of time in administration will be neither popular nor practical. Both teachers and pupils dislike tests that require a long time to do them and that require much writing to be done. In tests of this kind much energy is wasted in recording the answers after the problems have been solved mentally or after the correct answers have been determined to questions. Furthermore, the labor in scoring is so great that too much of the teacher's time must be spent in reading the papers. It isn't always easy to design a test that will satisfy these conditions, but the popularity of the test will depend very largely on these two points. The mechanical phases of test making are as important from the standpoint of administration as are the thought phases. Some of the most scientifically constructed tests we have, from the standpoint of accurately measuring the educational processes and products, are so mechanically clumsy that teachers dislike to use them. It may even be desirable that something be sacrificed in accuracy in order to increase the practicality of the tests. This is safe, however, only in so far as the net gain is greater than if most of the emphasis were placed on accuracy. The point may be illustrated from an example in the business world. From the standpoint

of accuracy the grocer might obtain a pair of scales that would measure sugar to one-one-hundredth of an ounce; but, practically, the net gain for both customer and grocer would be less than if he measured much less accurately.

CHAPTER VII

SCORING THE TESTS AND TREATMENT OF THE MEASURES

In this chapter the problems incident to the scoring of tests and the distribution of the measures after the tests are scored will be presented. While the teacher desires to know what each individual pupil is able to do on a test she also wants to know how the pupils stand as a class and how the class ranks when measured by established standards.

Problems of Scoring.—The problems of scoring are many and varied. The particular design of the test is very largely determined by the way the scores are to be obtained. In Chapter V we noted the problems incident to test making in a general way. In order to bring out more clearly the problems that must be met and solved in designing a test and especially to make clear the influence that scoring has on the general design of the test, the actual problems which confronted the makers of the Gregory-Spencer Geography Test¹ will be presented.

Example in the Development of a Geography Test.—The first problem that confronted the makers of this test was the question as to whether the test should be strictly *diagnostic*, covering quite thoroughly some specific field of geography, or a general test covering in a general way the entire field? The decision of this question was made only after a great deal of research work had been done

¹ Designed by C. A. Gregory, Professor of School Administration, University of Oregon, and Peter L. Spencer, Instructor in the University High School, University of Oregon. Published by the Bureau of Educational Research, University of Oregon.

attempting to find out what the best writers in the field of geography considered the purpose of geography to be. It was then necessary to consult the standard texts in geography in order to learn what subject-matter they contained and how the subject-matter was divided, and to make an estimate from the amount of space given to the various topics as to their relative values. There is probably no subject in the curriculum that lacks focus in presentation more than the subject of geography. The field is so broad and it contains so many phases where emphasis might be placed, that after one has exhausted every scientific means available, he must still rely in part on his personal judgment as to what constitutes the proper subject-matter for the test.

The question as to whether the test should be general or diagnostic was arbitrarily decided by making it, to a limited degree, diagnostic. Since the kind, amount, and order of presentation of geographic material are determined to a very large degree by the textbooks used, it would be folly to design a test covering subject-matter of a different kind. In the design of the test, therefore, the subject-matter *actually being taught*, rather than the subject-matter that *ought* to be taught, furnished the material for the test.

It is not the business of those designing a test to say whether the things taught are the things that ought to be taught. The purpose of the test is to determine how well, or to what degree, the children know and can do the things they are being taught. There are many lines of cleavage in the subject-matter of geography, and it was necessary to make divisions of some kind. It is possible to divide the field into *physical*, *political*, and *commercial* geography and make the tests with this division in mind. Or we might think of the subject-matter being divided into *causal geography*, *place geography*, *mathematical geography*, *map*

study, etc. Even granting that the test was to be made according to any one of these divisions, since the subject-matter in each division is so great that it could not possibly be covered by one, or even two or three tests, it was necessary to decide what particular bit of subject-matter should supply the material for the test.

State Examination Questions Examined.—It was thought that the questions prepared by state departments that are to serve as state examinations in geography might give some guidance in the preparation of the test. A letter was accordingly addressed to each state asking for lists of geography questions that the state departments prepare for the seventh and eighth grades. Forty-seven states replied and twenty-three prepared such questions. Some of the states sent questions dating back five years. The total number of questions thus received was about 1,300. After eliminating the local questions such as, "Bound your state or county," "Name the counties of your state," etc., the questions were compiled according to continents, countries, industries, etc. One of the striking characteristics of the questions is the fact that they lack focus. They are very widely scattered.

Another source of information was a representative group of courses of study, which were consulted for the purpose of finding out what phases of geography were there emphasized.

The Divisions Chosen.—Applying the information derived from the above sources, supplemented by our personal judgments, it was decided to divide the subject-matter for the test into the following divisions: (1) *Place and fact geography*, which test the pupil's knowledge as to where important cities, rivers, and seas are, and also his knowledge as to the distinguishing characteristics of a large number of cities. (2) *Causal geography*, which attempts to discover the pupil's power to reason from cause to effect.

This phase of the test was divided into two parts, one pertaining exclusively to the United States, and the other pertaining to the world as a whole. (3) *Commercial geography*. (4) *Political geography*.

The Selection of the Cities to Be Used in the Test.—Having decided that place and fact geography should constitute two of the major parts of the test, the next problem was to find out what cities, rivers, seas, etc., should be located, and what facts concerning them should be called for in the tests. In other words, since all the important cities could not be located in a test of this kind, and all the facts concerning the cities thus located could not be presented in the test, the problem of city selection was one that called for some method of choosing the most important cities. This was done in the following way. In the summer of 1910 Professor Whitbeck conducted a geographical seminar in Cornell University and had in his class about 75 teachers, principals, and superintendents representing 21 states. This class was divided up into committees and a continent was assigned to each committee with the instructions that the committee was to find the most important cities in them. They were to find cities that were so important that an American school teacher should teach their location rather accurately. They were also to teach why they were important and for what they stand in world affairs. It was agreed that a city to be included in the list must stand for more than one important thing. Lyons, for example, though it is the leading silk-making city of the world, has nothing else of importance that an American school boy needs to know. Hence it was not included.

These committees decided upon the lists of cities that should be taught and passed them over to a committee of the faculty on geography to be passed upon. This committee consisted of Professor R. H. Whitbeck; Professor Ralph S. Tarr, Cornell; Professor Albert T. Brigham, Col-

gate; Professor Charles McMurry; Philip Emerson, of Lynn, Mass.; and George D. Hubbard of Ohio State University.

Two-thirds of the cities listed by the first committee failed to pass the faculty committee. Any city of the United States that received two or more votes from the faculty committee was retained. No foreign city was retained unless it received at least three of the six votes of the faculty committee. The cities selected, together with the number of votes each received from the faculty committee, are given below:

UNITED STATES—25 Cities

New York, 6	Washington, 6
Chicago, 6	Denver, 6
Philadelphia, 6	Louisville, 2
St. Louis, 6	Minneapolis-St. Paul, 6
Boston, 6	Kansas City, 2
Baltimore, 2	Indianapolis, 2
Cleveland, 3	Duluth-Superior, 5
Buffalo, 3	Salt Lake, 3
Pittsburg, 6	Puget Sound Cities, 4
San Francisco, 6	Scranton-Wilkes-Barre, 3
Cincinnati, 2	Galveston, 4
New Orleans, 6	Lowell, 3
Milwaukee, 2	

FOREIGN COUNTRIES

Europe—16

London, 6	Naples, 6
Edinburgh, 6	Athens, 6
Glasgow, 6	Constantinople, 6
Madrid, 4	St. Petersburg, 6
Berlin, 6	Paris, 6
Hamburg, 6	Marseilles, 3
Vienna, 4	Venice, 4
Rome, 6	Liverpool-Manchester, 6

Asia—8

Bombay, 6	Hong-Kong, 5
Calcutta, 6	Jerusalem, 6
Canton, 5	Tokio-Yokohama, 6
Pekin-Tien-tsin, 6	Mecca-Medina, 3

Western Continent Exclusive of United States

Montreal, 6	Buenos Ayres, 6
Quebec, 5	Havana, 6
Rio Janeiro, 6	Mexico, 6

Africa, Australia, and Islands of Sea

Cairo, 5	Sidney, 3
Cape Town, 5	Manila, 6
Johannesburg, 3	Batavia, 3
Melbourne, 4	Honolulu, 6

The material having been selected and the amount having been determined upon, the next step was the actual drafting of the questions and statements that entered into the test. If the test is to be entirely objective, the personal equation of the teacher or other individual grading it must be entirely eliminated. If, therefore, pupils were allowed to frame their own answers to the various parts of the test it would call for an evaluation of these answers by the one scoring the papers. That is, it would rest on the judgment of the teacher whether the answer given by pupil *A* was of the same value as that given by pupil *B*. For instance, if the question, "Why is it dry east of the Rocky Mountains?" were asked there would be a variety of answers with varying degrees of merit, and since each answer might contain an element of truth, the one scoring the papers would be called upon to evaluate these answers. In order to eliminate this difficulty and also to simplify and lessen the labor in scoring the papers, the answers were framed by those making the test and the pupil was called upon

to put a cross after the best answer given. The following statements taken from the test, together with the instructions for giving them illustrate the point.

Below are given several facts about the United States. Three causes are suggested for each fact. Read the statements carefully, then place a cross (X) before the cause which you think best explains the fact.

1. The plains directly east of the Rocky Mountains are dry because:
 - (a) Few trees grow on them.
 - (b) The winds lose their moisture before they get to them.
 - (c) The land slopes eastward.
2. A large number of the people of Pennsylvania are engaged in the manufacture of iron and steel products, because:
 - (a) They have no lumber with which to build.
 - (b) Pennsylvania has many great iron mines.
 - (c) Pennsylvania has much coal with which to smelt the iron ore.
3. Seattle is farther north than Chicago, yet it has a milder climate because:
 - (a) Seattle is protected by the Rocky Mountains.
 - (b) Seattle is protected by heavy forests.
 - (c) The ocean modifies the winds which blow over it.
4. Cattle are raised on the Great Western Plains and are fattened and prepared for market on the prairie lands farther east, because:
 - (a) The market is in the east and the prairies produce much grain.
 - (b) It is warmer on the prairies and they afford better protection.
 - (c) The eastern people are richer and can afford to buy them.
5. New York City is called the Gateway to America, because:
 - (a) It is easy to get to the interior through this port.
 - (b) It is the largest city in America.
 - (c) It is on a navigable river.

Advantages of Tests Thus Designed.—There are at least five advantages of tests designed like those cited above:

1. *Time is saved for the pupil.*—The ideal test in a subject such as geography is one in which the pupil may cover the maximum amount of subject-matter in a minimum amount of time, and in which practically all the time may be spent in determining the proper answer or disposition of the parts of the test with a minimum amount of time in writing out or recording the answer thus formulated. In the first part of the test cited above, for instance, it is much easier simply to put a cross before the correct statement than to express the thought in the proper language and take the time to write it down on paper.

2. *Personal equation is eliminated.*—Each pupil is given a fair and unbiased evaluation of his ability in the subject-matter in question. If he has the cross before the correct answers his score will be as high as that of any other pupil in the class. The personal equation of the teacher is entirely eliminated. The pupil may look over his paper and know that the score given him is correct. The bonds of friendship between him and his teacher are thus strengthened, otherwise, there is, many times, a strained relation between pupil and teacher when the pupil thinks the teacher has not properly evaluated his paper.

3. *Much time is saved in scoring.*—If pupils were allowed to formulate their own answers, each word in the test would have to be read by the one scoring the papers, thus increasing the amount of labor many times. By the new method and with the aid of a key in the hands of one scoring the papers, the labor is reduced to the minimum and much of the drudgery is eliminated.

4. *More ground may be covered by a test thus designed.*—A pupil may cover four or five times as much ground by a test thus designed in the time allotted, and it may be covered more thoroughly. He is less fatigued, because

the labor of writing is eliminated, and he does not dislike a test of this kind because most of the drudgery has been removed and he knows he will get an unbiased evaluation of his work when it is completed.

5. *Pupils must review the whole field in preparing for the examination.*—Since pupils know that the entire field cannot be covered in an examination as now given, they spend considerable time trying to determine *what questions will probably be asked*, and spend their time on them instead of reviewing the entire field. The old alibi that in reviewing for the examination the pupil studied every part of the work but that covered by the examination would no longer apply, since the whole field would be covered.

Determination of the Scores.—It is extremely difficult for some teachers to believe that such a test as the one suggested above does anything more than give the highest score to the *luckiest guesser*. They say *it is a game of chance* that cannot possibly give a correct estimate of the pupil's school achievements. In spite of this prejudice, *chance is fatally exact*, and it is on this principle that the scoring may be done with an assurance that the real knowledge of the pupil may be determined in a test of this kind or, at least, as accurately as by the old method.

If one were to take a hundred pennies and toss them into the air one hundred times and count the number that fell heads up and the number that fell tails up, he would find that in the 10,000 tossed almost exactly half of them would fall heads up.

If, instead of there being three statements to choose from in giving the correct reasons in the above test, there had been only two, the chance would be analogous to tossing pennies. A student who knew absolutely nothing about the statements would get them right approximately 50 per cent of the time by *mere chance*. Then the question arises as to how the papers should be scored to take proper

cognizance of this game of chance. To do this we shall take a number of hypothetical cases to show that it is possible to make the proper evaluation in each case.

1. *When there is a choice between two answers.*—Let us suppose there are 20 parts to the test and the pupil actually knew 10 of them. Then he would mark his paper as follows: He would put a cross before the 10 he actually knew and guess at the other 10. By the laws of chance he would get 50 per cent of them right. Therefore, his paper would have 15 of the 20 parts of the test marked correctly. In order to make proper allowance for his chance scores we should subtract the number he marked incorrectly (5) from the number he marked right (15) in order to get his actual score, 10, because from our hypothesis we know that the number marked incorrectly would be half the number he guessed at. Therefore, his actual score would be 10, which, according to our hypothesis, is what he actually knew about the test.

Suppose again a pupil actually knows 16 of the 20 parts of a test. How would the scores show this fact? Since he actually knows 16 of the 20 parts of the test he would put the proper check marks before these 16 parts and guess at the other 4 parts. He would get 2 of these four parts right by the laws of chance. Therefore, his paper would have 18 answers marked correctly. Subtracting the number he had wrong (2) from the number he answered correctly (18), giving him a final score of 16, which is according to hypothesis.

2. *Where there is a choice among more than two answers.*—In the geography questions cited above, the student has a choice of one of three answers. By the laws of chance he would, therefore, get $33\frac{1}{3}$ per cent of them right even if he knew absolutely nothing about them. How would the papers be scored in a case of this kind? Let us take the case cited above and suppose there are 20 parts to

the test and that the student actually knows 11 of them. He would, therefore, put the proper check mark before the 11 that he knew and guess at the other nine, three of which he would get right by chance. His paper would, therefore, show 14 parts marked correctly. Now since he would get but one in three right by chance of the numbers he guessed at, we know that the number marked incorrectly would be twice as large as the number he marked correctly. Therefore, we would subtract one-half the number he marked incorrectly (3) from the total number marked correctly (14), thus leaving his final score 11, which is according to hypothesis.

If the number of answers to choose from was four instead of three and a student actually knew, let us say, 16 of the parts, his final score would be determined as follows: He would guess at four of them and get 25 per cent, or one, of them, right by chance. Therefore, his paper would have 17 parts marked correctly. Since he marked only one in four right by chance the number marked incorrectly would be three times as large as the number marked right by chance. Therefore, one-third of the number marked wrong, or one, would be the number to be subtracted from his total score in order to get his actual score.

Some Objections to Tests of This Kind.—It may be argued that as much information cannot be obtained by this method as the old method because a child told to discuss a certain topic would present knowledge that would take a dozen or more questions to bring out by the new method. There seems to be much truth in this criticism. It may be, however, that enough additional questions may be asked by the new method to offset the apparent loss in giving up the old system.

It is also argued that an examination of this kind is more or less superficial and does not involve, to any great extent, the higher thought processes such as reasoning,

imagination, etc. It would seem, however, that these processes must go on the same as by the other method. The only difference is that the new method does not impose the mechanical operation of actually writing out the answers. Some claim that the choice of words and the drill in sentence formation aids thought and that this is all lost when the answers are ready made. They claim that abstractions, comparisons, and reasoning do not take place to the same degree in the new method as in the old. It seems to the writer that these criticisms are not well founded.

Effect of Incorrect Statements Being Placed before the Student.—Another criticism is that because incorrect statements are placed before the children the psychological influence is bad, the false statements being taken as truths. In answer to this criticism it may be said that the facts do not warrant the statement that this condition prevails to an appreciable degree. Moreover the child would bring into consciousness the right and wrong answers in forming his judgments by the old method of giving examinations.

It is true that this method does not show where the reasoning goes wrong or ceases altogether; but it does save students the agony and perspiration necessary to perpetrate an answer like the following cited by McCall.² A student was asked the following question in a recent course in educational measurements: "Which three of the tests described by Whipple do you think would be of most service in an elementary school, if your school had a psychologist to apply them?" The answer was:

The tests described by Whipple embrace most of the difficulties that would be embraced in problems of classroom instruction. I

² "A New Kind of School Examination," *Journal of Educational Research*. Vol. 2, pp. 33-46.

think his tests embrace a great variety of methods of approach and it seems difficult for me to think of just three to whom the presence of a psychologist in a school would give help. I would think it would be tests in which knowledge of the workings of a child's mind and its growth and development would be most apparent since those not particularly trained might focus on others not of this kind. I feel it would be unwise to specifically mention just three when the number is so great which would fulfill all these requirements. Every teacher to be a psychologist would help all classroom measurement work of whatever kind greatly, I know since we cannot know of the influence of a test upon which any group except by the mental reaction produced.

In further support of the true-false examination it is maintained that it will promote a better feeling between the teachers and pupils; pupils will no longer strive to tell what they do not know; it eliminates the personal equation and makes a more pleasant atmosphere in general.

The Values to Be Assigned to the Scores.—In the foregoing discussion of scoring, the test questions have been *difficulty questions*, the problem being, "How hard a question can the pupil answer?" The answers were either *right* or *wrong* and each part was given an arbitrary value of 1. It is evident that some of these questions are more difficult than others. The question, therefore, arises as to whether or not the scores should be weighted so that credit may be given according to the relative difficulty of the various parts of the test. The defense of this method is that there is such a high correlation between the scores when the questions are weighted and when they are given an arbitrary value of 1 that the additional accuracy is not worth what it costs to get it.

Charters found a similar condition in making his language tests and did not weight his scores for that reason.

Some experimentation was done in the School of Education, University of Oregon, upon the correlation between the scores made where each part of the test is weighted

according to its difficulty and where the parts were given an arbitrary value of 1. The data were taken from the Douglass Standard Diagnostic Tests for Elementary Algebra, the Monroe Standardized Silent Reading Tests, and the Kansas Silent Reading Tests devised by Dr. F. J. Kelly. From random samples taken from papers in each of these tests, the correlation between the weighted scores and the scores where each part was given an arbitrary value of 1 was, in each case, taken as 0.9 or above.

General Problem of Weighting Scores.—In test making, however, if one wishes to weight the scores it may be done by determining the relative difficulty of each part of the test. This is usually done on the principle that the larger the number of children missing a part, the more difficult that part is and the larger score it should have. The following illustration from spelling will make the point clear. In giving the ordinary spelling examinations or tests the usual method of scoring is to mark the papers on a basis of 100 per cent. If there were 20 words, each word spelled correctly is given an arbitrary value of 5. It might so happen that two of the words in the test were "home" and "separate." The grading of the papers might show that 98 per cent of the fifth grade in a certain city system were able to spell the word "home" correctly whereas only 40 per cent were able to spell the word "separate"; yet, by the old method of marking the papers, a child would be given 5 per cent towards his final grade if he spelled the word "home" correctly and the same amount if he spelled the word "separate" correctly. It is evident that since the spelling difficulty of the latter is greater than that of the former, it should receive a higher score. If this is done it is called weighting the word on a basis of its spelling difficulty, which is probably the most scientific method of weighting the various questions or parts of tests in the tool subjects. The question as to

how much weight to give to a particular part of a test is generally determined in one of two ways.

1. *By the teacher's judgment.*—This is the usual method followed by teachers in giving the ordinary examination. If the scoring is to be done by the per cent method the teacher may arbitrarily say that question number 1 is worth 8 per cent and question number 2 is worth 12 per cent, and so on. She may weight the questions in this way on a basis of difficulty, her estimate being that question 2 is one-and-a-half times as difficult as question 1, or she may assign these weights, not because question 2 is more difficult than question 1, but because she thinks it is of more importance for social or other reasons that the children should know question 2. This method is rarely followed in test making.

2. *By weighting the parts according to the distribution of abilities as shown by the normal frequency curve.*—In this case the per cent of pupils missing each part is determined and these per cent values are transmuted into units of standard deviation (sigma) or probable error (P.E.). There are many methods that may be followed in doing this. One of the most common is perhaps to assign an arbitrary value of 1 to the easiest word or question in the test. This is determined as indicated above by finding the number of pupils who are able to answer it as compared with the rest of the questions in the test. The procedure may be illustrated as follows: Suppose the easiest word in a spelling test is spelled by 90 per cent of the pupils and the next word in order of difficulty is spelled by 80 per cent of the pupils. What should be the weighting assigned to these two words? If we desire to measure all the words in terms of the easiest word we may assign an arbitrary value of 1 to the first word.

Tables have been prepared so that it is possible to convert percentile scores into units of standard deviation and

weight them according to a normal distribution. For instance, in the above example a word missed by 10 per cent of the pupils is given a value of 1.73 (approximately), and one missed by 20 per cent is given a value of 2.16 (approximately). Therefore, the relative weights assigned to the two words are to each other as 1.73 is to 2.16. If we give the first word an arbitrary value of 1, then the weight or value of the second word would be 1.25.³ There is no particular reason why the easiest word or question should be given an arbitrary value of 1 other than the fact that some fractions may be avoided by this procedure.

Accumulation Scores and Scores of Greatest Difficulty.

—There are two general ways of determining the final scores of pupils. One method is to give each question or part a weighted value and let the sum of the scores of the various parts constitute the final score of the pupil. This is the method used in the Kansas Silent Reading Tests devised by Dr. Kelly, the Monroe Reading Tests, the Douglass Algebra Tests, and many others. The final score given a pupil is the accumulated values or weights given to each part. This procedure is followed also where the parts are not weighted but each part is given an arbitrary value of 1 as in the Curtis arithmetic tests. Here each problem is given a value of 1 and a pupil's score is the sum of the problems solved correctly.

The other method of determining the final score of a pupil is determined by the weighted value of the most difficult problem a pupil can solve. The Woody Arithmetic Tests illustrate this method of scoring. The final score given to a pupil is not determined by finding the sum of the weighted values of all the problems, but is simply the weighted value of the most difficult problem solved.

³ See Harold O. Rugg, *Statistical Method Applied to Education*, pp. 392-395.

The same principle is followed in the Thorndike Handwriting Scale. The score of the pupil is the highest quality reached and not the accumulation of the quality values ranging from the lowest to the highest quality reached by the examinee.

BIBLIOGRAPHY

1. BURGESS, MAY AYRES, *The Measurement of Silent Reading* (Department of Education, Russell Sage Foundation, 1921).
2. BURT, C., *The Distribution and Relations of Educational Abilities* (King and Son, London).
3. COURTIS, S. A., *The Gary Public Schools; Measurement of Classroom Products* (General Education Board, New York, 1919).
4. GRAY, C. T., *A Score Card for Measuring Handwriting*, Bulletin No. 17 (The University of Texas, 1915).
5. HAGGERTY, M. E., *The Intelligence Examination* (World Book Co.).
6. HAGGERTY, M. E., "Recent Developments in Measuring Human Capacities," *Journal of Educational Research*, Vol. 3, April, 1921.
7. HEALY, W. O., *The Individual Delinquent* (Little, Brown & Co., 1915).
8. HOLLINGWORTH, LETA S., *Vocational Psychology* (D. Appleton & Co., 1916).
9. "Intelligence and its Measurements; A Symposium," *Journal of Educational Psychology*, Vol. 12, March and April, 1921.
10. JUDD, CHARLES H., *Measuring the Work of the Public Schools*, Cleveland Educational Survey (Russell Sage Foundation, New York, 1916).
11. LINK, H. C., *Employment Psychology* (The Macmillan Co., 1916).
12. MCCALL, C. A., "A New Kind of School Examination," *Journal of Educational Research*, Vol. 1, pp. 33-46.
13. MCCALL, C. A., "Proposed Uniform Method of Scale Construction," *Teachers College Record*, Vol. 20, No. 1, January, 1921, pp. 31-51.
14. MÜNSTERBERG, HUGO, *Psychology and Industrial Efficiency* (Houghton Mifflin Co., 1913).
15. *National Intelligence Tests*, prepared by Haggerty, Terman, Thorndike, Whipple and Yerkes (World Book Co.).

16. National Society for the Study of Education, the various *Yearbooks* (Public School Publishing Co., Bloomington, Ill.).
17. Otis *Group Intelligence Scale*, designed by Dr. Arthur S. Otis (World Book Co.).
18. PINTNER, RUDOLF, and ANDERSON, MARGARET, M., *The Picture Completion Test*.
19. PINTNER, RUDOLF, and PATTERSON, DONALD, *A Scale of Performance Tests* (D. Appleton & Co., 1917).
20. ROSSOLIMO, "Mental Profiles; A Quantitative Method of Expressing Psychological Processes in Normal and Pathological Cases," *Journal of Experimental Pedagogy*, Vol. 1.
21. RUGG, HAROLD O., "Scientific Method in the Reconstruction of Ninth-Grade Mathematics," *Supplementary Educational Monographs*, Vol. II, No. 1, 1918.
22. RUSK, ROBERT R., *Experimental Education* (Longmans, Green & Co., 1919).
23. *Terman Group Tests of Mental Ability*, designed by Lewis M. Terman (World Book Co.).
24. TERMAN, LEWIS M., *The Measurement of Intelligence* (Houghton Mifflin Co., 1916).
25. YERKES, BRIDGES, and HARDWICK, *A Point Scale for Measuring Mental Ability* (Warwick and York, 1915).

CHAPTER VIII

THE MEASUREMENTS OF EDUCATIONAL PROCESSES AND PRODUCTS IN FIVE FIELDS OF SCHOOL WORK

In the latter part of Chapter II the entire field of tests and measurements was arbitrarily divided into seven divisions. The amount of work done in each division seemed to warrant such an arbitrary classification. It should not be inferred that the divisions made are the only divisions into which the field of measurements could be divided. It does seem, however, that such a classification is quite exclusive and there is comparatively little overlapping in these fields. Enough work has been done in each of them to give a great mass of data which are beginning to throw a great deal of light on the processes and products of education. In the last five chapters we have discussed and criticized the measurements of intelligence and the measurements of school achievements. In this chapter we shall discuss, very briefly, the other five fields, not with an idea of treating any one of them exhaustively but simply to call attention to the work that is being done along the lines mentioned in Chapter II. The fields are: (1) the measurements of the materials of instruction; (2) the measurements of the physical growth of school children; (3) the measurements of the money cost of education; (4) the measurements of school buildings; and (5) the measurements of retardation, acceleration, and elimination.

With the possible exception of the fourth category, sufficient facts and data are extant to form the basis of a large volume, and, in some cases, many volumes, on each of

these fields of measurement. It was also pointed out in Chapter II that each of these fields may be subdivided into smaller divisions and that some are combined to form new fields. Tests in vocational guidance, for instance, may involve measures of intelligence, measures of school achievements, and physical measurements. Other measures are similarly combined to form new and definite fields. Without further discussion of the broad general field of measurements we shall deal more specifically with the various divisions as outlined.

I. MEASUREMENTS OF THE MATERIALS OF INSTRUCTION

It is only within the last decade that any considerable work has been done in measuring the materials of instruction. To some it has seemed like educational pedantry to count the words in a spelling book, or to score a textbook of any kind in order to get an analytic conception of its contents. On the other hand, the folly of presenting material year after year with little knowledge of its content and with no quantitative conception as to the relative amounts of the various elements that compose it has led many students to an intensive study of the materials of instruction.

Determination of a Spelling Vocabulary.—It was by studies of this kind that a spelling vocabulary of school children and also of adults was determined. Each individual of school age, or, at least, after he has passed the first two or three years of his school life, has four vocabularies: (1) reading, (2) speaking, (3) hearing, and (4) spelling, or writing. Spelling being a tool, there would be no need for learning to spell words unless an individual would use these words when he writes. The problem then is to discover what words an individual uses when he writes. This can be done only by taking the written com-

positions of those who write and analyzing them in order to determine the words used and their frequency. These compositions may include business letters, friendship letters, compositions written in school, the composition of a daily newspaper, and many other types of material.

Since it is not possible to teach all the words that one may use when he writes, the best that can be done is to determine the words that occur in greatest frequency and teach them as a minimum word list. Many perplexing problems present themselves in determining the proper word lists, such as: How may one find out what words should be taught? Ayres¹ made up his list of 1,000 words by combining the results of four studies in spelling. One study was made by Rev. J. Knowles of London, England, and was published in pamphlet form under the title, "The London Point System of Reading for the Blind" (1904). In making this study the author took passages from the Bible and other literature, containing in all 100,000 words, and from this list took the 353 words of the greatest frequency.

The second study was made by R. C. Eldridge of Niagara Falls. Eldridge made an analysis of 250 different articles taken from four issues of four Sunday newspapers published in the city of Buffalo. He found that they contained a total vocabulary of 6,002 different words and 43,989 running words.

The third study was made by Ayres and published by the Russell Sage Foundation in a monograph entitled, *The Spelling Vocabularies of Personal and Business Letters*. This study consisted of a tabulation of 23,629 words from 2,000 short letters written by 2,000 people. The total vocabulary used was found to consist of 2,001 different

¹ *A Measuring Scale for Ability in Spelling* (Division of Education, Russell Sage Foundation, 1915).

words. The number of appearances of each was reported in the monograph.

The fourth study was made by Cook and O'Shea and the results presented in a book entitled *The Child and His Spelling*, published in 1914. This study consists of a tabulation of approximately 200,000 words taken from the family correspondence of 13 adults. The total vocabulary was found to be 5,200 different words.

The list of 1,000 commonest words in the Ayres Scale was finally selected from these four studies by finding the frequency with which each word appeared in the four studies, weighting that frequency according to the size of the base of which it was a part, adding the four frequencies thus obtained, and finding their average.

Anderson² attempted to determine the spelling vocabulary by analyzing the words found in 5,000 letters gathered by school children from the various sections of the state of Iowa and incorporating the words of greatest frequency into a minimum spelling list. His list contains approximately 5,000 words.

Another problem that must be solved is the question as to how great a frequency a word must have for each 100,000 running words before it is incorporated in the minimum spelling list. This must be decided arbitrarily. For instance, suppose a word occurs but once in 100,000 running words. Is that frequency sufficient to justify its being taught in the elementary school? Or, should the writer be referred to the dictionary to find how to spell a word whose frequency is but 1 in 100,000 running words? It is probably safe to say that a word with a frequency of less than 3 in 100,000 running words should not occur in a speller for the elementary school.

² *The Determination of a Spelling Vocabulary Based upon Written Correspondence* (Ph.D. thesis, University of Iowa, 1917).

Having determined the list of words that should be taught in the elementary school, the next problem is to determine the *order* in which they should be taught. This again involves measurements in order to properly grade the words. Several factors enter into this problem. Words might be graded on a basis of their spelling difficulty, the easiest words being assigned to the lower grades, and the more difficult ones to the upper grades. Or again they might be graded on the basis of use. Children in the lower grades do not use quite the same vocabulary as children in the upper grades. Jones³ found that when children, from grades 2 to 8 inclusive, were asked to write compositions, the average vocabularies from grade to grade were as follows:

Grade	Number of Words
2	521
3	908
4	1,235
5	1,489
6	1,710
7	1,926
8	2,135

If words were selected exclusively on a basis of use, the words used by children in the second grade should constitute the spelling vocabulary for that grade and additional words used by children in the third grade, together with those left over from the second grade, would constitute the words for the third grade, and so on. Use, frequency, and difficulty are the chief factors which must determine the grading of the words in the subject of spelling. Extensive measurements have been made in all three factors and the grade in which a word is now presented

³ *Concrete Investigation of the Material of English Spelling* (University of South Dakota), p. 23.

is no longer a mere matter of opinion but is the result of scientific investigation.

A Study of the Reading and Spelling Vocabularies of Books Used in the First Three Grades.—Perhaps the most thorough and complete study of the words that children of the first three grades are called upon to read and spell has been made under the direction of the author by Miss Ruth Chase.

The problem originated in the following way: The Text-book Commission of the state of Oregon met and adopted a list of books to be used by children in the first three grades of the elementary school for a period of six years. The books, with the number of pages in each, are given below:

1. *Beacon Primer*, 124 pages
2. *Natural Primer*, 122 pages
3. *Beacon First Reader*, 130 pages
4. *Natural First Reader*, 136 pages
5. *Natural Second Reader*, 256 pages
6. *Natural Third Reader*, 304 pages
7. *Hamilton's Essentials of Arithmetic*, 124 pages
8. *New World Speller*, Reading Vocabulary
9. *New World Speller*, Spelling Vocabulary, 124 pages⁴

The Problem.—The study was made to determine: (1) the number of different words a child would be called upon to read, spell, and understand, to meet the minimum requirements of the state course of study for the state of Oregon; (2) how rapidly a child's reading and spelling vocabularies grow, assuming that the work is taken up in the order indicated by the state course of study; (3) the frequency of the words used; (4) the correlations between the books used, that is, the number of words common

⁴ The *New World Speller* was divided into two parts for purposes of scoring. * The eighth book mentioned above deals with the reading material a child must master in this book, and the ninth book deals exclusively with the words he must learn to spell.

to all the books; (5) the number of running words in the series.

Terms Used in the Study.—*Different words* means the number of separate words used in each book. If the proper name "Mary Jones" were used, it is listed as two words. The plural and singular of a word were counted as two words.

There were 503 different words used in the *Natural Primer*, although 265 of these were not counted as new words because they had been used in the *Beacon Primer*.

New words. A word used for the first time is called a new word. In the *Beacon Primer* each word was counted as a new word because this book was the first one used in the series. The sum of the new words used in each book is the total number of different words used in all the books.

Used words are words which have occurred in an earlier book of the series. The *Natural Second Reader* contains 1,770 different words, of which 861 are *new* and 909 are *used words*.

Running words means the number of times all the words occur. For example, if 12 different words are found and the sum of their frequencies is 49, the number of *running words* is 49.

The Results.—Table III shows the distribution of words in each book. The numbers at the heads of the columns refer to the readers named above:

TABLE III.—DISTRIBUTION AND SIZE OF VOCABULARIES

	1	2	3	4	5	6	7	8	9
Number of running words.....	7,007	6,315	8,455	8,878	25,332	34,855	12,574	2,705	6,767
Number of different words.....	743	503	804	849	1,770	3,304	1,282	450	1,453
Number of used words.....	0	255	404	584	909	1,416	753	392	1,240
Number of new words.....	743	238	395	265	861	1,888	529	58	213
Increase in vocabulary from book to book..	743	981	1,376	1,641	2,502	4,390	4,919	4,977	5,190

The study reveals the following facts:

(a) 4,977 different words constitute the child's minimum reading vocabulary, to which are added 213 words he must learn to spell which are not found in his reading vocabulary, making a total of 5,190 different words.

(b) 106,121 is the number of running words.

(c) 289 words, or 5.8 per cent of the child's reading vocabulary, are used 75,591 times and constitute 71.2 per cent of the running words in the reading vocabulary.

(d) 13 words occur 27,458 times.

(e) 10 words occur 24,520 times.

(f) 1,470 words, or 29.9 per cent of the words, occur but once.

(g) 1,453 words were listed in the speller 6,767 times, or an average of 4.6 times.

(h) 3,218 words occur 29,060 times, an average of 9 times.

(i) 289 words occur 75,591 times, an average of 261 times each.

(j) 213 words were found in the speller which were not found in the readers.

TABLE IV.—THIRTEEN WORDS OF GREATEST FREQUENCY FOUND IN THE READERS COMPARED WITH THE AYRES' LIST

Rank in Readers	Frequency in Readers	Rank in Ayres' List
1. the	7,927	1. the
2. and	3,142	2. and
3. to	2,441	3. of
4. a	2,336	4. to
5. he	1,790	5. I
6. of	1,714	6. a
7. I	1,629	7. in
8. in	1,473	8. that
9. you	1,385	9. you
10. was	1,343	10. for
11. said	1,069	11. it
12. it	942	12. was
13. is	927	13. is
	27,458	

^a *A Measuring Scale for Ability in Spelling* (Division of Education, Russell Sage Foundation, 1915).

It is interesting to note that the 13 most frequent words found in this study are identical with those of the Ayres list except one. This list contains the word "said," which is not found in the Ayres list; and the word "that" is found in the Ayres list but not in the thirteen most frequent words in this list.

The Contents of Three American Histories.—Another study made under the direction of the author that illustrates measurements of the materials of instruction is the scoring of three textbooks in American history. In June, 1919, the Textbook Commission for the state of Oregon readopted Mace's *School History of the United States* for a period of six years. The book had been in use for a number of years in Oregon and was generally conceded to be a satisfactory text in United States history. From the fact that it was to be the official text in American history for the next six years it was thought that its usefulness might be increased if its contents were scored in reference to some of the most salient features that are being discussed in the reorganization of history courses in the grades. With this thought in mind, a class of advanced and graduate students taking a course in the elementary curriculum with the author at the University of Oregon undertook as a special problem to score the book in reference to four salient features. In order to provide some standards for evaluation and comparison, two of its competitors were scored with it. These books were *History of the American People* by Beard and Bagley, and *History of the United States* by Gordy. The scoring was done in reference to the following points:

1. *Names of places*—scored under the following headings:

- (a) Names of continents and countries
- (b) Names of states and territories
- (c) Names of rivers
- (d) Names of cities and towns
- (e) Names used in connection with military events

- (f) A miscellaneous list which did not fit under any of the above headings
2. *Names of men*—scored under the following headings:
- (a) Explorers and discoverers
 - (b) Rulers, presidents, and governors
 - (c) Names mentioned in connection with industry and invention
 - (d) Names of statesmen
 - (e) Names mentioned in reference to military matters
3. *Dates*
4. *Amount of space devoted to political, social and economic, and military matters; also the number of pages devoted to pictures, maps, and illustrations.*

TABLE V.—NAMES OF THE TEN COUNTRIES OF GREATEST FREQUENCY

<i>Mace</i>		<i>Beard and Bagley</i>		<i>Gordy</i>	
Name	Frequency	Name	Frequency	Name	Frequency
England.....	250	United States.	333	England.....	220
United States.	129	England.....	122	United States.	158
France.....	91	France.....	93	France.....	62
Spain.....	77	Great Britain.	82	Spain.....	58
Canada.....	44	Spain.....	68	Mexico.....	27
Mexico.....	42	Germany....	60	Canada.....	18
India.....	16	Mexico.....	35	Cuba.....	15
Great Britain.	16	Russia.....	24	China.....	11
Holland.....	14	China.....	21	Japan.....	11
Alaska.....	12	Cuba.....	19	East India...	10

TABLE VI.—SUMMARY OF COUNTRIES AND CONTINENTS MENTIONED

	<i>Mace</i>	<i>Beard and Bagley</i>	<i>Gordy</i>
Total number of countries mentioned.....	42	55	48
Number mentioned but once.....	14	15	16
Total number of mentions in each text.	785	1,069	687
Number of continents mentioned.....	6	6	5
Frequency of mention of continents.....	286	224	161

TABLE VII.—NAMES OF THE TEN STATES AND TERRITORIES OF GREATEST FREQUENCY

<i>Mace</i>		<i>Beard and Bagley</i>		<i>Gordy</i>	
Name	Frequency	Name	Frequency	Name	Frequency
Virginia.....	138	Virginia.....	86	Virginia.....	76
New York.....	75	Massachusetts	62	Massachusetts	57
Massachusetts	72	Pennsylvania..	58	New York....	51
Pennsylvania..	58	New York....	54	South Carolina	42
South Carolina	44	South Carolina	43	Georgia.....	25
Maryland....	42	Ohio.....	39	Louisiana.....	24
Texas.....	36	Texas.....	38	Connecticut...	24
Tennessee....	36	California....	38	North Carolina	21
Kentucky....	36	Kentucky....	34	Florida.....	19
New Jersey...	33	Illinois.....	33	Pennsylvania..	19

Mace mentions 51 states and territories with a total frequency of 1,066; 49 of them he mentions two or more times. Beard and Bagley mention 51 states and territories with a total frequency of 1,085, all of which are mentioned two or more times. Gordy mentions 48 states and territories with a frequency of 572, eight of which are mentioned but once.

Table VIII records the number of rivers mentioned in the three texts. Mace mentions 56 rivers with a total frequency of 256, of which 32 appear but once. Beard and Bagley mention 35 with a total frequency of 142, of which 19 appear but once. Gordy mentions 43 with a total frequency of 219, of which 22 appear but once.

Table IX records the number of cities appearing in the three texts with their frequencies. Mace mentions 153 cities with a frequency of 720, 77 of which are mentioned but once, and 26 are mentioned twice. Beard and Bagley mention 186 cities with a total frequency of 682, of which

TABLE VIII.—NAMES OF THE TEN RIVERS OF GREATEST FREQUENCY

<i>Mace</i>		<i>Beard and Bagley</i>		<i>Gordy</i>	
Name	Fre- quency	Name	Fre- quency	Name	Fre- quency
Mississippi....	60	Mississippi...	47	Mississippi...	60
Hudson.....	25	Ohio.....	21	Hudson.....	34
Ohio.....	24	Hudson.....	9	Ohio.....	28
Potomac.....	24	Delaware....	7	Shenandoah..	10
Delaware....	13	Potomac.....	7	St. Lawrence..	8
St. Lawrence..	11	Columbia....	5	Mohawk.....	8
Connecticut...	8	Missouri....	4	Potomac.....	5
Rio Grande...	8	Rio Grande...	4	Connecticut..	5
James.....	6	Arkansas....	3	Tennessee....	4
Niagara.....	5	St. Lawrence.	3	Delaware....	4

105 appear but once, and 30 are mentioned twice. Gordy mentions 140 with a total frequency of 529, of which 72 appear but once, and 16 appear twice.

TABLE IX.—NAMES OF THE TEN CITIES OF GREATEST FREQUENCY

<i>Mace</i>		<i>Beard and Bagley</i>		<i>Gordy</i>	
Name	Fre- quency	Name	Fre- quency	Name	Fre- quency
New York....	65	New York...	68	New York...	41
Philadelphia..	60	Philadelphia..	48	Boston.....	34
Boston.....	59	Boston.....	42	Philadelphia..	32
Washington...	45	New Orleans..	26	Washington...	31
Charleston....	31	Chicago.....	25	New Orleans..	19
Chicago.....	25	Washington..	22	Richmond....	14
London.....	22	Charleston...	18	Charleston...	14
New Orleans..	21	St. Louis....	17	Hartford....	12
Albany.....	17	Pittsburgh...	16	Baltimore....	10
St. Louis.....	17	Buffalo.....	12	Albany.....	10

Table X records the places mentioned in reference to military events. Mace mentions 172 places with a frequency of 433, 81 of which occur but once, and 34 are mentioned twice. Beard and Bagley mention 127 places with a frequency of 260, of which 70 occur but once, and 24 twice. Gordy mentions 184 places with a frequency of 560, of which 85 occur but once, and 31 are mentioned twice.

TABLE X.—NAMES OF THE TEN PLACES OF GREATEST FREQUENCY APPEARING IN CONNECTION WITH MILITARY AFFAIRS

<i>Mace</i>		<i>Beard and Bagley</i>		<i>Gordy</i>	
Name	Fre- quency	Name	Fre- quency	Name	Fre- quency
Richmond.....	18	Concord.....	9	Richmond....	17
Vicksburg.....	14	Richmond....	9	Mississippi	
Chattanooga...	9	Boston.....	8	River.....	16
Gettysburg....	9	New York....	8	New York....	16
Yorktown.....	9	Washington,		Hudson River.	16
Concord.....	8	D. C.....	6	Virginia.....	16
Ft. Sumter....	8	Virginia.....	6	Washington,	
Quebec.....	8	Lexington....	6	D. C.....	15
Trenton.....	8	Gettysburg...	5	Philadelphia...	14
Bunker Hill....	7	Santiago.....	5	Boston.....	12
		Philadelphia...	5	South Carolina	11

Table XI gives a miscellaneous list of places which did not fit into any of the above classifications. Mace mentions 140 places with a total frequency of 484, of which 88 places are mentioned but once, 17 twice, and 5 three times. Beard and Bagley mention 184 places with a total frequency of 696, 112 of which are mentioned once, 25 twice, and 18 three times. Gordy mentions 136 places with a total frequency of 569, 72 of which are mentioned once, 23 twice, and 18 three times.

MEASUREMENTS IN OTHER FIELDS 245

TABLE XI.—MISCELLANEOUS LIST OF NAMES OF PLACES MENTIONED

<i>Mace</i>		<i>Beard and Bagley</i>		<i>Gordy</i>	
Name	Fre- quency	Name	Fre- quency	Name	Fre- quency
New England..	85	South.....	94	South.....	114
Cuba.....	33	North.....	71	North.....	85
Atlantic Ocean..	20	New England..	49	New England..	57
West Indies....	19	West.....	41	Allegheny Mts.	18
Confederacy....	16	Pacific Ocean..	27	West.....	16
N. Netherlands.	15	East.....	24	Atlantic Ocean.	11
Philippines....	14	Confederacy...	16	Lake Champlain	9
District of Co-		Atlantic Ocean.	12	Pacific Ocean..	8
lumbia.....	13	Mississippi		Lake Erie.....	8
New Amsterdam	9	Valley.....	10	New World...	7
East Indies....	8	Great Lakes...	9		

GRAND TOTAL OF PLACES MENTIONED

Mace	480
Beard and Bagley.....	507
Gordy	284

To this total should be added a number of places connected with military events, which would bring the grand total up somewhat higher.

From the foregoing tables one is impressed with the vast number of geographical facts a child in the seventh and eighth grades must know to read these histories intelligently.

If mere frequency of mention is in any way indicative of the importance of a place, the rivers, cities, places connected with military events, countries, etc., may thus be evaluated.

In comparing the frequencies of mentions of the cities, rivers, countries, states, places mentioned in reference to

military events, and even those in the miscellaneous lists, one is struck by the great similarity throughout the three texts. The first ten places receiving the greatest number of mentions in one text are practically the same as those in the others.

If this is a fair evaluation, the teacher may be justified in taking, say, the ten places of highest frequency from each of the six tables, making sixty places in all, for more or less intensive geographical study.

Table XII is a summary of five tables of men mentioned in the three textbooks.

Space would not permit more than a summary of each of these tables.

TABLE XII.—NAMES OF MEN MENTIONED

	<i>Mace</i>	<i>Beard and Bagley</i>	<i>Gordy</i>
Explorers and discoverers.....	47	34	30
Rulers, presidents and governors.....	87	69	66
Names mentioned in connection with indus- try and invention.....	11	40	8
Names of statesmen.....	29	*	44
Names mentioned in reference to military matters.....	167	76	104

* Data not extant.

Table XIII gives the dates that had a frequency of five or more in each of the three texts, the distribution of dates in the study, "Possible Defects in the Present Content of American History as Taught in the Schools," reported by Horn in the *Sixteenth Yearbook* of the National Society for the Study of Education, Part I, and also the study made by Bagley in the *Fourteenth Yearbook* in the same series.

MEASUREMENTS IN OTHER FIELDS 247

TABLE XIII.—DATES APPEARING FIVE OR MORE TIMES

<i>Mace</i>		<i>Beard and Bagley</i>		<i>Gordy</i>		Frequency of Dates Found by Horn		Dates Ranked According to Importance in Bagley's Study	
Date	Fre- quency	Date	Fre- quency	Date	Fre- quency	Date	Fre- quency	Date	Rank
1862	16	1860	39	1862	17	1900	114	1776	1
1812	13	1850	25	1812	14	1890	84	1492	2
				1900	14				
1864	12	1917	20	1837	13	1850	56	1607	3
1860-63	11	1913	19	1861	12	1870	53	1789	4
1776	9	1861	17	1850	11	1825	52	1620	5
1865	9					1893	52		
1837-50-87	8	1862	16	1763	10	1902	43	1803	6
				1820	10				
1814	7	1865	15	1807-60-64	9	1880	40	1861	7
1846	7								
1750-63-77	6	1916	14	1863-65-93	8	1910	38	1787	8
		1812	14						
1844-73	6								
1636-39-64-82	5	1816	13	1781	7	1869	33	1863	9
				1828-35-90	7				
1754-78-94	5								
1811-16-21-20	5								
1825-28-47-72	5								
1876-97	5								
		1776	12	1660	6	1860	28	1820	10
		1864-70	12	1754-79-87-89	6				
		1910	12	1832	6				
		1900	11	1609-89	5	1913	27	1812	11
				1814-15-30	5				
				1836-46-54	5				
				1857-73-77	5				
				1895-99	5				
				1801	5				
		1848-63	10			1848-82	26	1765	12
		1763	9			1871	22	1783	13
		1828-25-67	9						
		1880-96	9						
		1912	9						
		1836-40-95	8			1840	20	1865	14
		1908	8						
		1787	7			1857	14	1857	15
		1800-10-11	7						
		1844-72-90	7						
		1815-20-33	6			1825	12	1854	16
		1837-46-47	6						
		1858-69-79	6						
		1887-88-92	6						
		1893	6						
		1909-11-14	6						
		1769-92	5			1830	11	1775	17
		1806-01-14	5						
		1819-22-41	5						
		1854-66-68	5						
		1871-77-78	5						
		1897-98	5						
		1901-04-05-07	5			1790	8	1781	18
						1800-03	8		
						1819-20	7	1823	19
						1794	4	1846	20

Table XIII is read as follows: The date 1862 occurred 16 times in Mace's *History*; the date 1860 occurred 39 times in Beard and Bagley; the date 1862 occurred 17 times in Gordy; the dates 1844 and 1873 occurred six times in Mace; the dates 1848 and 1863 each occurred five times in Beard and Bagley, and so on.

The following facts in reference to the dates in the three history texts are significant. Mace's *History* contains 326 different dates; Beard and Bagley, 235; Gordy, 206. It is interesting to note that the ranks of the history dates found in these three history texts do not agree with those found by Dr. Horn in the study of "Thirty-Eight Modern Crucial Problems" in the above-mentioned study, or with the dates selected by specialists in American history and reported by Bagley in the *Fourteenth Yearbook*. The most important date found in Mace's *History*, as far as frequency is concerned, is 1862; in Beard and Bagley, 1860; in Gordy, 1862; in Horn's study, it is 1900; and in Bagley's study, 1776.

TABLE XIV.—AMOUNT OF SPACE DEVOTED TO POLITICAL, SOCIAL AND ECONOMIC, AND MILITARY PHASES OF AMERICAN HISTORY

Type of Material	<i>Mace</i>		<i>Beard and Bagley</i>		<i>Gordy</i>	
	Pages	Per Cent of Book	Pages	Per Cent of Book	Pages	Per Cent of Book
Political movements	144.65	30.13	240.50	37.81	148.44	33.96
Social and economic movements	150.00	31.28	155.37	24.43	63.22	14.46
Military movements	98.40	20.50	39.50	6.21	66.22	15.50
Maps and illustrations . . .	83.75	17.40	97.95	15.40	80.33	18.38

The above percentages do not include the bibliographies,

review questions, and other repeated materials in the texts. One significant thing about these texts is the amount of space devoted to pictures and maps. Some of the books have approximately 15 per cent of their space devoted to pictures. Since the picture cost is more than the cost of a regular printed page it brings the approximate cost of the pictures up to 20 per cent of the entire cost of the book. When one-fifth of the manufacturing cost of a book is the cost of reproducing pictures, one should be pretty sure that the pictures are valuable and will be generally used.

If the contents of these books are compared with the type of material discussed by Horn, cited above, it will be seen that in the former relatively much less space is devoted to social and economic movements. However, it was not the purpose of the studies to determine the relative amount of space that ought to be given to these various phases of American history but rather to determine the amount of space that was actually being given in these three texts. The same thing may be said in reference to the other phases scored.

II. THE MEASUREMENTS OF THE PHYSICAL GROWTH OF SCHOOL CHILDREN

In the last analysis, education may be reduced to the process of producing, directing, and preventing changes in human beings. This process has to do with the physical changes as well as the mental. It is a platitude to say that the physical and physiological traits of the individual primarily condition his total nature; yet the exact relation that exists between the physical and mental development of an individual is not known.

Much work has been done on the measurement of the physical growth of children, largely because some scientists

have believed that measurements of this kind would in some way furnish a key to mental development. Hundreds of thousands of school children have been measured to determine their height, weight, vital capacity, reaction time, and other physical traits. Norms for these traits for each year of a child's life, until he gets far into the 'teens, have been established. It would seem that there should be a somewhat definite relation existing between the development of the nervous system and the mental maturity of an individual; that frail undersized children should be taught somewhat differently from the strong and robust; and that courses of study should be reorganized to take cognizance of the development of physical traits of school children. Yet in spite of the hundreds of thousands of measurements that have been made on these physical traits, little change that takes cognizance of the growth and degree of maturity of these physical traits has been made in the organization of school programs. Some time in the near future the facts that have been obtained by the physical measurements of school children may be extremely valuable in the teaching process. To date, however, their pedagogical value has been small. To state the problem concretely, how should the method of the recitation or the type of subject-matter presented to children who are exceptionally well developed physically differ from that presented to those children who are scarcely normal as far as physical development is concerned? It very frequently happens that those who are physically inferior exceed those whose physical development is far above the normal, and the work done by those physically inferior apparently does not injure them. Of course, many valuable facts relative to growth have been discovered that are beginning to be utilized in health work among children.

III. THE MEASUREMENTS OF THE MONEY COST OF EDUCATION

So many measurements have been made in this field, and the literature is so familiar, that a bare mention of it is sufficient here. We have made careful records of the cost of school buildings, school apparatus, teachers' salaries, and other expenses incident to teaching. Recently we have worked out norms for supervision, teaching, janitor service, and the cost per clock-hour to teach the various subjects. A school official may now compare the distribution of school funds in his city with that in other cities and determine whether he is paying too much or too little for supervision, janitor service, or any other phase of the school work.

Intensive studies have been made in school costs and accounting of school work. This phase of measurements is of interest primarily to the department of school supervision rather than to the average room teacher.

IV. THE MEASUREMENTS OF SCHOOL BUILDINGS

The score card for the measurement and standardization of school buildings came as a result of the movement for greater efficiency in education. The first one made its appearance in 1916 and was made under the direction of Dr. George D. Strayer of Columbia University. A new *Score Card for City School Buildings* was made by Strayer and Engelhardt in 1920. This card has two special purposes: (1) *The scoring of school buildings in the light of a school building program to be developed by a city.* (2) *The checking of plans for new school buildings.*⁶ This score card has grown out of the experience in evaluating more than 1,000 school buildings and the study of school

⁶ *Score Card for City School Buildings*, Teachers College Bulletin No. 10, Jan., 1920.

building standards. Like many of the scales used in the measurements of school achievements, it is a score card based on the combined or median judgments of experts. A hall one foot too narrow is not wrong in the same sense that an addition problem is wrong; but it is not the most efficient precisely because it is the combined opinion of competent judges that it should be one foot wider.

When one considers the amount of money that is spent in the erection of school buildings, he wonders why an attempt to standardize school buildings was not made sooner. Individuals judging buildings not infrequently think mainly in terms of two or three, or possibly a half dozen, elements that seem to them of primary importance and often neglect other parts of the building that are of equal importance. The score card is designed to include as nearly as possible all these details that go to make up a perfect school building. In assigning weights to the various elements the judgment method is resorted to. The scoring is done on a basis of 1,000 points; that is, a perfect building scores 1,000. The principal divisions, or headings, for scoring city school buildings with the weights assigned to each heading are given below:

1. Site	125
(a) Location	55
(b) Drainage	30
(c) Size and form.....	40
2. Building	165
(a) Placement	25
(b) Gross structure	60
(c) Internal structure	80
3. Service systems	280
(a) Heating and ventilating	70
(b) Fire protection system	65
(c) Cleaning system	20
(d) Artificial lighting system	20
(e) Electric service system	15
(f) Water supply system	30

MEASUREMENTS IN OTHER FIELDS 253

(g) Toilet system	50	
(h) Mechanical service system	10	
4. Classrooms		290
(a) Location and connection	35	
(b) Construction and finish	95	
(c) Illumination	85	
(d) Cloakrooms and wardrobes	25	
(e) Equipment	50	
5. Special rooms		140
(a) Large rooms for general use.....	65	
(b) Rooms for school officials.....	35	
(c) Other special service rooms.....	40	

Each of the above headings have a number of subheadings which go into detail as to the various parts of the buildings.

The score card for the measurement of school buildings is meeting with approval everywhere. Large cities that have voted down school bonds for improvement and repairs have voted twice the sum called for when the buildings have been scored and the exact defects made known to the public.

It seems evident that a score card of this nature will eventually become a standard for the erection of most school buildings.

V. THE MEASUREMENTS OF RETARDATION, ACCELERATION, AND ELIMINATION

Measurements of retardation, acceleration, and elimination were among the first to be made in the recent general movement in educational measurements for greater school efficiency. The movement may be said to have started in earnest in 1904 when Dr. William H. Maxwell, city superintendent of the schools of New York, showed in his annual report that 39 per cent of the pupils in the elementary grades were above the normal age for the grades they were in.

This startling revelation caused school officials to check up their school systems to see if similar conditions prevailed elsewhere. From that time to the present those in charge of the schools in every state in the Union have been measuring the amount of retardation in their schools and attempting to assign causes for the retardation found.

It was the importance of this problem with its bearing on the question of the adaptation of the school to the needs of the child, and the almost complete lack of definite information bearing on the question, that impelled the Russell Sage Foundation to undertake in 1907 an investigation of "some phases of the adaptability of the school and its grades to children."⁷ The investigators were interested, not in the individual subnormal, or in a typical child, but rather in that large class, varying with local conditions from 5 to 75 per cent of all the children in our schools, who were older than they should be for the grades they were in. Data gathered for this study seemed to warrant the statement that at least 6,000,000, or 33 per cent, of the pupils in the public schools were retarded.

Thirteen per cent more retardation was found among boys than among girls. The percentage of girls who completed the common school course was 17 per cent greater than the percentage of boys. Studies of this kind brought out the fact that the schools as then organized were better fitted for the needs of girls than they were for the needs of boys.

It was also found that there was a high correlation between retardation and elimination. In those schools where the pupils were greatly retarded, a large majority did not remain to finish the course.

The report of the Commissioner of Education in 1907 shows the distribution of children through the grades in

⁷ Leonard P. Ayres, *Laggards in Our Schools* (Russell Sage Foundation, 1907), p. 2.

386 cities of 8,000 population and above. From this report it was shown that for every 1,000 pupils entering the first grade, the second grade would have 723, the third grade, 692, the fourth grade, 640, the fifth grade, 552, the sixth grade, 462, the seventh grade, 368, the eighth grade, 263, the first year of high school, 189, the second year of high school, 123, the third of year high school, 81, and the fourth year of high school, 56.

Of course conditions have improved since 1907 as a result of the studies that have been made on retardation and elimination. The following study made by Ayres is probably typical of the progress that is being made.

A little more than ten years ago the Department of Education of the Russell Sage Foundation made a coöperative study of 200,000 school children in 29 city school systems to determine their progress through the grades.⁸ In the spring of 1920 the same schools were asked to repeat their earlier study in order that some estimate could be made of the progress, if any, attained during the interim. Fifteen cities repeated the tests using the same procedure and the same record blanks that had been used in the first test. The 15 cities repeating the test had 83,283 children in 1911 and 111,680 in 1920.

In working out the age-grade tables a pupil who was seven years old and in the first grade was considered of normal age and one year was added for each advancing grade. A pupil could thus be classified both according to his age and his progress. In regard to age, he was either younger than normal, normal, or older than normal. With respect to his progress he was either slow, normal, or rapid. Since both age and progress were recorded and there were three groups in each classification, each child could be

⁸ Leonard P. Ayres, "The Increasing Efficiency of Our City School Systems," *Elementary School Journal*, Vol. 21, Feb., 1921, pp. 416-423.

assigned to any one of nine different classes. Table XV shows the classification for each 100 pupils in 1911.

TABLE XV.—SCHOOL CHILDREN BY YOUNG, NORMAL, AND OLD, AND BY RAPID, NORMAL, AND SLOW GROUPS, FIFTEEN CITIES, 1911
(After Ayres)

	Young	Normal	Old	Total
Rapid.....	6	3	2	11
Normal.....	20	21	11	52
Slow.....	2	9	26	37
Total.....	28	33	39	100

Table XV is read as follows: Six children were younger than normal for their grades and had progressed faster than normal. The 21 who appear in the second column were of normal age and made normal progress, and so on.

Table XVI shows the conditions in 1920, computed on a basis of 100 pupils.

TABLE XVI.—SCHOOL CHILDREN BY YOUNG, NORMAL, AND OLD, AND BY RAPID, NORMAL, AND SLOW GROUPS, FIFTEEN CITIES, 1920
(After Ayres)

	Young	Normal	Old	Total
Rapid.....	10	2	1	13
Normal.....	28	23	7	58
Slow.....	2	9	18	29
Total.....	40	34	26	100

The data of Table XVI show that conditions in 1920 were better than in 1911. The children who were both young and making more than normally rapid progress had increased from six in each 100 to ten. Those in the center of the table, of normal age and making normal

progress, had increased from 21 per cent to 23 per cent. The most important change is that in the figures in the lower right-hand corner which shows that the unfortunate misfits who were over age and making slow progress had diminished from 26 per cent to 18 per cent.

During the nine years the percentage of over-age children had fallen from 39 to 26 and the proportion of slow pupils from 37 in each 100 to 29 in each 100. These improvements are large and important. They represent educational economy, financial saving, and human conservation.⁹

When the survey movement in the schools started about 1912, the questions of retardation and elimination received a great deal of attention. A great deal of work has been done in this field by Ayres, Strayer, Thorndike, and others. The field has been rather completely covered in state and city surveys and other official reports.

The question of the bright child, the accelerated child, is just now receiving much attention. Educators are beginning to realize that perhaps the bright child has suffered from our ill-adapted school system rather than the dull one. Besides, it is the bright child who will eventually become a leader in society, a moulder of public opinion, and hence the child who will yield the largest income on the investment made. Schools are accordingly being re-organized to take special cognizance of the bright child. Large sums of money are being donated to educators for special work in this field. Large cities are holding summer sessions to make it possible for the bright child to gain a half grade or even a grade during six or eight weeks in the summer. Special classes are being provided for him in most of the larger school systems and in many of the more progressive smaller ones.

⁹ *Ibid.*, pp. 418-419.

All of this means measurements. The bright children are selected by determining their general intelligence and school achievements. It is not too much to prophesy, perhaps, that the chief reorganization of the school will be along lines which will take cognizance of the individual differences among children.

BIBLIOGRAPHY

1. AYRES, LEONARD P., "The Increasing Efficiency of Our City School Systems," *Elementary School Journal*, Vol. 21, Feb., 1921, pp. 416-423.
2. ANDERSON, *The Determination of a Spelling Vocabulary Based Upon Written Correspondence* (Ph. D. thesis, University of Iowa, 1917).
3. BAGLEY, W. C., "The Determination of Minimum Essentials in Elementary Geography and History," National Society for the Study of Education, *Fourteenth Yearbook*, Part I, pp. 131-146.
4. BURGESS, MAY AYRES, *The Measurement of Silent Reading* (Russell Sage Foundation, New York, 1920).
5. CHAPMAN, J. C., *Scientific Measurement of Classroom Products* (Silver, Burdett & Co., Boston, 1917).
6. COURTIS, S. A., *The Gary Public Schools, Measurement of Classroom Products* (General Education Board, New York, 1919).
7. GREGORY, C. A., "The Reading Vocabularies of Third-Grade Children," *Journal of Educational Research*, Vol. 5, 1922.
8. GREGORY, C. A., and SPENCER, PETER L., "A Geography Test for the Sixth, Seventh and Eighth Grades," *School and Society*, Vol. 15, 1922.
9. JUDD, CHARLES, H., *Measuring the Work of the Public Schools*, Cleveland Educational Survey (Russell Sage Foundation, New York, 1916).
10. JONES, W. FRANKLIN, *Concrete Investigation of the Materials of English Spelling* (University of South Dakota).
11. MONROE, WALTER S., *Measuring the Results of Teaching* (Houghton Mifflin Co., 1918).
12. MONROE, W. S., DE VOSS, J. C., and KELLY, F. J., *Educational Tests and Measurements* (Houghton Mifflin Co., 1917).
13. PINTNER, RUDOLF, and PATERSON, DONALD, *A Scale of Performance Tests* (D. Appleton & Co., 1917).

14. RICE, J. M., "The Futility of the Spelling Grind," *Forum*, Vol. 23, pp. 163-172, 409-419.

15. STARCH, DANIEL, *Educational Measurements* (The Macmillan Co., 1917).

16. STRAYER, GEORGE, and ENGELHARDT, N. L., *Score Card for City School Buildings* (Teachers College, Columbia University, 1920).

17. "Standard Tests for the Measurement of the Efficiency of Schools and School Systems," National Society for the Study of Education, Part I, *Fifteenth Yearbook* (Public School Publishing Co., Bloomington, Ill., 1916).

18. Terman, Lewis M., *The Measurement of Intelligence* (Houghton Mifflin Co., 1916).

19. "The Measurement of Educational Products," National Society for the Study of Education, Part II, *Seventeenth Yearbook* (Public School Publishing Co., 1918).

20. Wilson, G. M., and HOKE, J. KREMER, *How to Measure* (The Macmillan Co., 1921).

21. WHITBECK, R. H., "Where Shall We Lay the Emphasis in Teaching Geography?" *Education*, Vol. 31, pp. 108-116.

22. WOODY, CLIFFORD, *Measurement of Some Achievements in Arithmetic*, Teachers College Contributions to Education, No. 80, 1920.

23. Various Conferences on Educational Measurements, Indiana University Bulletins (University of Indiana, Bloomington, Ind.).

CHAPTER IX

EDUCATIONAL STATISTICS, GENERAL STATEMENT

The mastery of one more set of tools is necessary before the educator may consider himself fully equipped to speak intelligently about his educational processes and products. Twenty-seven million school children now sit at the feet of 750,000 teachers to receive instruction. The annual cost of operating this great institution is well beyond the billion dollar mark. In a great many school systems more than 100,000 school children are under the direction of a single educator. The school systems have grown so large and there are so many individual characteristics and variations that affect the various groups that it is impossible, without the aid of additional tools by which we may compare, contrast, and weigh one tendency with another, to get an idea of the general movement of the group as a whole.

The human mind is so constituted that it cannot image and comprehend a large number of distinct impressions at any one time. For example, he would be a man with exceptional mnemonic power who could listen to the reading of two lists of grades of 100 each that were made by students taught by two different methods in the subject of addition and tell which method was the better as shown by the scores made by the pupils in each group.

The power needed to detect the movements of groups as a whole, when the movements of the individuals within the group are many and varied, is to be found in the elements of statistical methods.

Use of Statistics in Other Fields.—The educator is not a pioneer in this field, but is simply following the lead of the biologist, the economist, the sociologist, the natural scientist, and others. The importance of statistical methods applied to these sciences is rarely recognized. The whole doctrine of evolution and heredity rests in reality on a statistical basis. It is in this direction that the most important new work of a statistical nature is being done. Out of the great number of observations, such as the measurements of the height of a group of men, the *type* is found, that is, the average about which all the measurements are grouped according to some definite law. The problem is then to determine whether this type, or the groupings about it, change, and in what way. The differences found in successive generations form the data on which arguments as to evolution and development are founded. The same methods apply equally to fossil remains, to zoölogical species, and other organic forms. If this method were neglected many valuable arguments would lose their force and theories would be based on personal impressions of phenomena instead of on scientific measurements.

In certain sciences a higher degree of accuracy is required than is possible with a single measurement. For every measurement, however apparently absolute it may be, is a *relative thing*, made in terms of something else that is also fallible; that is to say, that is subject to variation. All scientific measurements, in other words, are made in terms of units theoretically invariable, but which are always practically applied by means of a mass of matter used to measure with, which, itself, is of necessity a more or less variable quantity; how variable, being again, in turn, a matter of determination. Therefore, many measurements are made of the same magnitude, and the average taken as the true amount. The astronomer, for instance,

uses statistical methods in getting the position of the heavenly bodies. The method of *least squares*¹ was introduced by him in locating the position of a star because he was anxious to choose the best of several slightly discrepant observations of the position of the star.

The Question of Error.—In all physical and biological observations the usual method is to take several measurements of the same quantity in order to get the most nearly accurate result obtainable as a measure of the thing in question. To all such measurements there enters the factor of *experimental error* due to a number of causes such as environmental conditions, the apparatus, or the observer himself.

The errors fall in two general classes: They are (1) the *constant errors*, which, in all measures of the same quantity, made with the same care, and under the same conditions, have the same magnitude, or, whose presence and magnitude are due to some fixed cause; and (2) the so-called *accidental errors* such as those due to fatigue, cold, nervousness, poor eyesight, or other temporary disability of the experimenter, or to the constitutional bias known as the *personal equation*. Some errors, of course, are simple mistakes, as in reading off the wrong figure, mistaking a 3 for a 5, for instance.

After a full investigation of the constant errors in all physical and biological measurements, the problem then remains of combining the observations so that the remaining accidental errors shall have the least probable effect upon the results, and it is to bring about this combination of observations that we employ the *method of least squares*.

Distribution of Measures about a Point of Central Tendency.—The averages obtained by different scientists

¹ Used to locate a position such that the sum of the squares of the distance from that point is the minimum.

from the same series of biological, sociological, and other observations are rarely identical. From such a group of measurements it is necessary to deduce the most probable estimate. It was early discovered that the measurements obtained by different scientists measuring the same thing showed a certain definite arrangement in accordance with which values at or near the average of all the measures were greatest in frequency; that positive errors were about as frequent as negative ones of the same magnitude; and, that large errors seldom occur. The center of balance about which the errors, or observations fall, is known as the *arithmetical mean*. In reference to the arithmetical mean as a measure of the most probable value of a series of measurements on the same thing, Merriman says:²

The most probable value of a quantity which is observed directly several times with equal care, is the arithmetical mean of the measurements. The average, or arithmetical mean, has always been accepted and used as the best rule for combining direct observations of equal precision, upon one and the same quantity . . . if the measurements be but two in number, the arithmetical mean is undoubtedly the most probable value; and, for a greater number, mankind, from the remotest antiquity, has been accustomed to regard it as such.

For example, out of ten discrepant results, it is impossible to ascertain the true value. What is the best representative of that value? Experience has shown the arithmetical mean to be the best, that is, the most representative value of a series of observations made under the same conditions, all being equally reliable. The arithmetical mean is so regarded, because it is a value, the deviations from which in the plus and minus directions, being equally probable, will cancel one another. Or again, if one hundred judges

² *A Textbook on the Method of Least Squares* (New York, 1913), p. 22.

were called upon to judge the length of a certain room which was actually 80 feet long, it would be found upon tabulating the data that the most of the guesses were relatively close to 80 feet, and as the guesses deviated farther and farther from 80 their number would grow consecutively less. It would be found also that the number of guesses that exceeded 80 were practically equal to those that were less than 80. In the cases just mentioned, and others of a similar nature, the mean simply represents a center of equilibrium, or center of gravity, as it were, of the variations in the given measurements. Having found that center of equilibrium—the arithmetical mean—the next problem is to ascertain what amount of swing or oscillation there may be on either side of this center. These are the variations, which correspond to the errors we have referred to, in the making of physical measurements. The questions arising relative to variability and measures of central tendency are treated more at length in subsequent chapters.

The graphic representation of the distribution of data in reference to a measure of central tendency as discussed above is known as a *normal* or *probability surface*, and the curve representing it is known as a *normal probability curve* discussed in a previous chapter.

Measures differing from the average as discussed above were thought of as being *in error* and this gave rise to the development of what has been called *theory of error*.

Considerable space has been given to an exposition of the significance of the theory of error in this general chapter on educational statistics because a correct idea of it is fundamental to the discussion that follows.

Educational Measurements Compared with Measurements in Other Fields.—We have thus far discussed the question of error where a large number of measurements were made of the same thing. We may now assume that

there is an analogy between a series of measurements of the same thing and a series of single measurements of each of a number of individuals alike in some important characteristics. In biology, in particular, this analogy was found to work very well but was less applicable to economic data. For this reason there have grown up two schools, the one adhering to the doctrine of the *theory of error* and the other rejecting it in the main. In the treatment of educational data it is found that educational measurements resemble those of biology in their structure, that is, that the theory of error applies.

As investigations in new fields of science were constantly being made and as the data became more complex, the need for better statistical methods became more imperative. Refined statistical inquiries could not be conducted by the crude and cumbersome machinery then in operation. As a result of this need the development of the pure theory of statistics has had a remarkable growth within the last two or three decades. Such men as Francis Edgeworth, August Meitzen, Francis Galton, Edward L. Thorndike, Karl Pearson, G. Udny Yule, and Charles B. Davenport have contributed in the field of biological statistics; and Arthur L. Bowley, Jacques Bertillon, R. H. Hooker, Thomas S. Adams, and Warren Persons have each aided in establishing statistical methods in the field of economics.

In dealing with large numbers descriptive of groups in any field it is found that special methods become necessary; methods that depend on the peculiar properties of large numbers; methods that are suitable for describing complex groups so they can be easily comprehended; methods for analyzing the accuracy of statements, for measuring the significance of differences, and for comparing one estimate with another. All of these fall within the scope of statistics. Without the aid of statistical methods, we simply have large numbers and groups of numbers from which

no logical deductions can be drawn. It must be borne in mind, however, that *statistical methods in themselves prove nothing*. The methods selected for use in a particular situation must agree with the logic and other non-quantitative facts of that situation. When thus used, statistical methods aid us in refining our thinking about complex masses of data and also in refining our methods of expression. Bowley says: "The proper function of statistics, indeed, is to enlarge individual experience."³ Because statistics measure only the numerical aspects of a phenomenon they should be brought into relation to the personal, political, æsthetic, and other non-quantitative considerations that may be of greater importance in deciding on a course of action.

Quantities Measured Indirectly.—Statisticians must very often content themselves with measuring, not the facts they wish, but some allied quantity, since it is frequently the case that the quantity about which knowledge is desired is not capable of numerical measurements. We cannot measure health, crime, or poverty, for instance. We can measure only death-rate, the number of convictions, or the number of persons who receive public relief. Many facts in other fields are thus measured.

Few important actions can be taken by a modern government or even a modern corporation without a statistical study of the conditions of the field in question. Statistical results are essential when judgments are to be formed on any question which involves numbers, quantities, or values; but they should always be used with discretion and care. "The most important function of statistics," says Bowley, "is evidence to show the relation of one group of phenomena to another."⁴ The information obtained is presumed intended as a guide for action.

³ *An Elementary Manual of Statistics*, p. 8.

⁴ *Ibid.*, p. 1.

Definition of Statistics.—Statistics has been defined as the science of averages. They render the meaning of masses of figures clear and comprehensible at a glance. They give a bird's-eye view of a situation involving a complex series with numerous cases in such a way that we get a picture of the series as a whole. They have to do with movements of groups as a whole. They refer to a large mass of facts, or data, that bear upon some human problem. This is one of the most important principles involved in statistical methods. The individual members of a group change rapidly while the whole group changes slowly. It is impossible, for instance, to follow or measure the motions of the separate atoms of a body, but comparatively easy to measure the motion and state the laws governing the movements of a body as a whole. When we wish to obtain a measurement of a group, peculiarities of individuals receive little attention. It is only when the same peculiarities are possessed by a considerable number of persons or things that they become of importance and are taken into account.

Statistics are numerical statements of facts in any department of inquiry placed in relation to each other; statistical methods are devices for abbreviating and classifying the statements and making clear the relations. Statistics are almost always comparative. They show the relative importance, the very thing an individual is most likely to misjudge. The absolute magnitude of a quantity is of little meaning until we have some similar quantity with which to compare it. The object of a statistical estimate of a complex group is to present an outline, to enable the mind to comprehend with a single effort the significance of the whole.

Statistics deal primarily with variable quantities. A *variable* is a quantity that, under the conditions imposed, may assume different values throughout a discussion. In

education, where we make numerous single measurements of different pupils grouped together on the basis of some common characteristic, each different value constitutes a value of the variable.⁵

Laws of Statistical Regularity.—One of the most valuable contributions of modern scientific statistics is that it has succeeded in giving us a sufficient picture of a group of objects without going through the laborious and expensive process of a complete enumeration of all the items in the group. Thus, it is by no means necessary, in ascertaining the average wage of American workingmen, to obtain data regarding each man at work. If certain typical instances are taken and properly averaged, the difference of this average from the true average wage of all workingmen is likely to be such a small quantity as to be, for all practical purposes, negligible.

In a similar way the anthropologist can discover the physical characteristics of a tribe or race by taking careful measurements of only a small minority of the whole. This is due to a law of nature formulated on the mathematical theory of probabilities, that a moderately large number of items chosen at random from among a very large group are almost sure, on the average, to have the characteristics of the larger group. Thus, if two persons blindfolded were to pick here and there 300 ears of corn each from a bin containing 1,000,000 ears, the average length of the ears picked by each person would be almost identical even though they varied considerably in length. It must not be inferred from the above that any number of samples, no matter how large, will give exactly the same results as would be obtained by the use of the entire mass of data. The probability of error diminishes con-

⁵ B. R. Buckingham, "Statistical Terms and Methods," National Society for the Study of Education, *Seventeenth Yearbook*, Part II, p. 115.

stantly as the number of items used increases. If, then, only a few sample items are used, the chance error is likely to be so large as to seriously vitiate the results; but as the number of samples chosen grows larger, the error diminishes until it eventually becomes negligible.

Methods of Statistics.—The quantitative study of education reveals two principal methods of treating measurements of human traits and other educational data. On the one hand, the observer may note only the *presence* or *absence* of an attribute or trait. For example, if 98 degrees Fahrenheit is considered a normal temperature for an adult human being, and if an individual who has a temperature higher than that is said to have a fever, then an observer may examine individuals to see whether or not this attribute, the fever, is present. The quantitative character in this case arises solely in the counting of the number of individuals who possess this attribute. The method by which we treat statistics collected in this way has been defined by Yule as the *statistics of attributes*.

On the other hand, the observer may want to know, not only as to the *presence* or *absence* of the attribute, but *how much* of the attribute is present. In the case cited above, he may want to know *how much* fever each individual has. This method of refining statistics and measuring the actual magnitude of variable attributes is known as the *statistics of variables*. This is the method usually employed in educational research. For example, we want to know, not only whether or not there is retardation in our schools, but how much retardation; not only how many pupils made grades above the passing marks, but how much above the passing marks; not simply that there is elimination from school, but how much elimination; and so on.

This method implies that the magnitude of the attribute or characteristic has been measured with reference to some *scale* made up of *units*.

Limitations of Statistics.—Statistics, while extremely useful to the investigator in almost every line of scientific inquiry, have limitations and shortcomings which cannot be overcome. Statistics deal largely with averages and these averages may be made up of individual items radically different from each other. In the average these irregularities are swallowed up. But statistics, from their very nature, cannot and never will be able to take into account individual cases. The difference between arithmetic and statistics is that the former attains exactness while the latter deals with estimates.

Standard of Accuracy.—While in the physical sciences very great accuracy of measurements is practicable, this is far from being true in the case of social phenomena. In this field a multitude of sources of error are ever present, many of which can be eliminated by no degree of care. Fortunately for the statistician, however, small errors are often negligible and in no way obstruct the solution of the given problem. Attempts to attain the greatest possible degree of accuracy are frequently merely waste of time. It might be possible to measure the customs revenue of the United States to the nearest cent, for instance, but for ordinary purposes of statistical comparison, such action is not only superfluous but positively confusing to the mind, in as much as the addition of extra figures directs the attention from the fundamental digits.

Compensating vs. Cumulative Errors.—The accuracy of the final results depends very largely on whether the errors are compensating or accumulative. If different people were to estimate the length of a given line the chances are that as many people would estimate it too long as too short. The errors in measuring a line made by a pair of chainmen, because of stretching the chain too tight or not taking up the slack sufficiently, would tend in the long run to offset one another. In cases of

this kind the errors are said to be *compensating*. On the other hand, if the chain used by the above-mentioned surveyors were too short, the longer the line measured the greater the error would be. The last case mentioned shows the effects of *accumulative errors*.

Discrete and Continuous Series.—Quantities to be measured may be in either a *discrete* or a *continuous* series. A *discrete series* is one with gaps. It is made up of a number of integers, as the number of words in a spelling test, or the number of children in a class. There are either 20 words in a spelling test, or 21, or some other integral number. There are never $20\frac{1}{2}$ or $20\frac{2}{3}$ words in the test; neither are there $19\frac{1}{2}$ children in the class. In each case the number is an integer.

A *continuous series* is one that does not contain gaps and is, in theory, capable of any degree of subdivision. Most mental traits and social facts belong to this series. In actual measurements, a given measure of a continuous series does not mean a single point on the scale, but a distance along the scale between two limits. For instance, when we say that an athlete runs 100 yards in $10\frac{5}{10}$ seconds, we do not mean that it was exactly $10\frac{5}{10}$ seconds but that it was at least $10\frac{5}{10}$ and less than $10\frac{6}{10}$. A more delicate recording instrument might have recorded the time as $10\frac{57}{100}$ seconds, or as $10\frac{574}{1000}$, and so on, depending on the delicacy of the instrument.

Undistributed Measures.—The fact that many of our marks and measures in education are indefinite and undistributed leads to a great many errors in educational statistics. A few illustrations will make this point clear. In a continuous series the measure zero, which should be a definite distance on the scale, a measure somewhere between two limits, is quite indefinite and very confusing when used as a point of reference in statistics. Unless the statistician defines what he means by zero, correct reckon-

ing in reference to it is impossible. Thus zero may mean from minus 0.5 to plus 0.5 or from 0 to 1 or some other measure on the scale. It, many times, means a distance on the scale from a point above *just not any* of the thing to be measured to an indefinite distance below. If ten boys were given an examination in arithmetic and the test consisted of ten problems, a boy who failed to solve any of the problems would be marked zero, and unless otherwise explained this mark would mean anything less than one to an unknown lower extreme. The boy who solved six problems is scored six; but a score of six unless otherwise explained means as much as six and less than seven. These points should be carefully noted to prevent confusion in finding medians and other measures of central tendency and variation in subsequent pages.

Rules for Tabulating Data.—The first thing that the statistical investigator must decide is the exact nature of the problem that he desires to solve. The first essential is to make the problem definite and clear-cut. The next problem is the arrangement of the data in a frequency distribution. At first thought it would seem to be one of the simplest things in the world to construct a frequency table for recording data; but the beginner who attempts to tabulate a complex group of figures will quickly discover that the simplicity of the operation is far more apparent than real. In fact, when a scientific tabulation has once been made, it is often found that a large share of the work of analysis is completed.

In beginning a tabulation the first question that arises is whether to put the figures in one or in several tables. A single table has the merit of completeness, and the data are thus brought into proximity. The table, however, if too large, becomes confusing to the eye and there is great difficulty in following the lines and columns at a glance. Each table should be a unit. Rarely should

one attempt to demonstrate in the same table several comparisons of different natures. Another matter to decide is whether the table shall show absolute figures, or percentages, or both. The number of separate headings, or columns, is a third query which must be answered. The more minute the subdivisions, the greater is the accuracy obtained. On the other hand, the multiplicity of headings prevents the proper emphasis being given to the main facts and tendencies shown by the statistics.

General Directions for Making a Scale and Curve-Plotting.—Scales and distribution tables are necessary in statistics for two reasons:

1. The object of the scale may be to present graphically a vivid picture of the general distribution of the facts relative to a given problem. One of the most common ways of representing these facts so that the eye will catch their general trend at a glance is to *plot the curve*. This is done in the following manner: A horizontal straight line is first drawn, and points are located at equal distances on this line. At the left end of the line a perpendicular is erected, and points are laid off in a similar manner on this line. The two series of points are called the scales. It is usual to call the point where the perpendicular line intersects the horizontal line the origin of coördinate, which is designated by *O*. The horizontal line is usually designated by *OX* and is called the *X*-axis. The vertical line is designated by *OY* and is called the *Y*-axis. Distances along the *X*-axis are spoken of as *X*-distances or *X*-coördinates, and distances along the *Y*-axis as *Y*-distances or *Y*-coördinates. In making a distribution table and plotting curves, lines may be drawn parallel to the *X*-axis through the points located on the *Y*-axis, and lines may be similarly drawn through the points on the *X*-axis parallel to the *Y*-axis, or, the curve may be plotted without drawing these additional lines. The curve is more

easily analyzed, however, if these additional lines are drawn.

The procedure in plotting a curve may be illustrated as follows:

Lay off on the X -axis distances equal to the magnitude of the part or trait measured, and, at the respective distances representing each magnitude, erect perpendiculars to the X -axis. Similarly plot the corresponding traits on



FIGURE VI. PLOTTING OF A CURVE

the Y -axis and erect perpendiculars. The intersections of the perpendiculars thus erected constitute the desired curve.

2. The second reason for making scales and distribution tables is to arrange the measures so as to facilitate computation.

Designating the Class Intervals.—Different methods are used in designating the class intervals. In some instances

the class interval is designated by its mid-point. If the heights of individuals are under consideration, for instance, and the steps are 60, 62, 64, 66, 68, etc., 60 may mean from 59 to 61, and 62 may mean from 61 to 63, etc. In the Courtis arithmetic tests the size of the step is 1 problem. It is not designated by the middle of the step, but by its lower limit. For example, 6 problems does not mean from 5.5 problems to 6.5 problems, but means from 6 to 7 problems, and 7 problems means from 7 to 8 problems, etc. That is, a pupil would get credit for only 7 problems even though the eighth one were almost completed.

When the limits of the steps are not clearly designated, the computation is difficult to follow. It is, therefore, a good policy for the beginner to state clearly what his class intervals are before he begins the computation.

In any class interval there may be many measures. The measures are spoken of as the *frequencies of the class intervals*, and the total frequency is the sum of all the class frequencies.

Analysis of Results.—The results disclosed by a distribution table are seldom fully revealed at a glance. Much is therefore added to the value of a table if it is accompanied by a written analysis which points out the principal conclusions which may be deduced therefrom, the possible errors involved, and the probable causes of the phenomena. The power to analyze a table, interpret the results correctly, and state the conclusions lucidly and succinctly is one of the characteristics indispensable in a good statistician.

In studying things of the same variety the work may usually be facilitated by dividing the items into classes. The simplest mode of classification is to group all the instances under two headings, the determining factors being whether they do, or do not, possess a given characteristic.

Thus we may classify people as sane or insane, workmen as employed or idle, flowers as white or colored, men as short or tall, and so on.

For some purposes this division by *dichotomy*, or cutting in two, may be most satisfactory but in many cases the difficulty arises that there is no distinct dividing line. Thus, it is impossible to say at just what point a man ceases to be short and becomes tall. It is therefore necessary to lay off arbitrarily a line of demarcation between the two classes. But if classes are to be thus arbitrarily established, it is often much more advantageous to set up a large number of them rather than only two. In practice this is usually done by dividing the whole group into classes of equal width. Thus, if the tallest trees in a group are 39 feet and the shortest 16, and it is desired to divide the entire group into five classes, the boundary lines would preferably be fixed on the round numbers 15, 20, 25, 30, 35, and 40. These boundary lines are known as *class limits*, and the distance between the two limits of any class is designed as a *class interval*. In the case cited above, 5 would be the class interval. A table formed by thus dividing the group into a number of smaller, more homogeneous classes, and indicating the number of items found in each class is known as a *frequency table*. The number of items falling within the given class constitutes the *size* of that class or the frequency.

The Need for Understanding Statistical Formulas.—The real scientist must know his tools. He owes it to the science and to the persons who may accept his results to be quite familiar with his tools. The blind application of formulas in statistics has been encouraged by the convenient manuals that are available and by the fact that the theory has been surrounded by intricate and involved mathematics so that the non-mathematical student had great difficulty in interpreting them. There is little doubt

that the failure of many books on statistical methods to set forth the fundamental principles involved in the treatment of statistical data has done much to hinder the progress in the use of statistics in education. The necessary mathematics is largely elementary arithmetic, and, with a few exceptions, there is no need for higher mathematics. Special effort has been made in this volume to present the fundamental principles of statistics simply and, as far as possible, in non-mathematical language. Practically all higher mathematics has been eliminated.

CHAPTER X

THE MEASUREMENTS OF CENTRAL TENDENCY OR AVERAGES

The previous chapters, dealing with the general nature and use of statistics, the methods of organizing materials in the form of frequency distributions, and the definitions and illustrations of statistical terms, have prepared us to take up the various methods of statistically treating the distributions thus organized. It will soon become clear that the organization of material into frequency distributions is but a preliminary step to a more concise description and representation of it by analyzing the size, number, and position of the various items that make up the distribution.

Many things about the nature of the distribution must be known before comparisons with other distributions may be scientifically made. The limitations of the various averages must be fully recognized. The amount of dispersion about the averages must be known, and the reliability of the measures determined.

It has been indicated in a previous chapter that we are primarily interested in comparative values rather than in absolute values. In fact, a thing cannot be evaluated until it is compared with something else. Statistical data cannot be compared until they are expressed in terms that properly represent the entire mass. There are three principal ways of describing statistical data in the form of frequency distributions, that is, there are three general types of measurements of statistical data. They are: (1) *measurements of central tendency, or averages*; (2)

measurements of dispersion or variability; and (3) measurements of correlation. The first of these will be discussed in the present chapter and the other two in the two succeeding chapters.

Averages.—The word average has a very indefinite meaning in common parlance. The public uses the term very loosely. Generally speaking, the word average as used by the public may mean one of two things: It may mean the most frequent measure in the group, "the general run," the typical measure, as the average clerk, the average teacher, the average size city, the average American; or, it may mean a different thing altogether, illustrated by what the farmer has in mind when he says that his hogs averaged 210 pounds, or his wheat averaged 14.7 bushels to the acre.

In the second interpretation of the average, it may be that not a single hog in the drove actually weighted 210 pounds, nor a single acre actually yielded 14.7 bushels of wheat; hence this average is very different from the other which is the typical measure in the series.

The first average illustrated above is more specifically known as the mode, and the latter as the arithmetic mean, or simply the mean. A discussion of the characteristics and limitations of these averages and also those of another average called the median will be taken up at some length.

The Arithmetic Mean.—The arithmetic mean may be defined as *the sum of all the measures in a distribution divided by the number of measures.* It is represented by the formula

$$M = \frac{\Sigma fm}{N}$$

where M represents the arithmetic mean of the distribution, Σ indicates that the products of fm are to be added, m = the value of any measure, f = the number, or fre-

quency, of the measure of a given value, and N the total number of measures. Table XVII illustrates the simple computation of the arithmetic mean.

TABLE XVII.—GRADES IN PER CENT MADE BY TEN STUDENTS IN FIRST-YEAR ALGEBRA
(Hypothetical Case)

Pupil	Grade m	Frequency f	Frequency \times Measures fm
A	84	1	84
B	91	1	91
C	68	1	68
D	79	1	79
E	95	1	95
F	93	1	93
G	87	1	87
H	85	1	85
I	90	1	90
J	92	1	92
			10 $\overline{)864} = \Sigma fm$
			86.4 = the mean

In the treatment of statistical data the distributions are rarely so simple as in the above illustration. Instead of having one student make a grade of 84, another 91, and so on, it would be more probable to have several students making grades of 84, 91, etc., ranging all the way from about 60 per cent up to 100 per cent.

Table XVIII is more nearly representative of the way data are found and treated statistically.

Table XVIII illustrates a distribution where each score is made by more than one pupil; that is, a score of 4 is made by three pupils, a score of 5 is made by four pupils, a score of 6 by 21 pupils and so on. Since there is more than one person making each of the various scores, the distribution is said to be *weighted*, and the arithmetic mean found from such a distribution is called a *weighted arithmetic mean*.

TABLE XVIII.—DISTRIBUTION OF SCORES ON 614 SAMPLES OF PENMANSHIP MADE BY CHILDREN IN THE THIRD GRADE IN THE SALT LAKE CITY PUBLIC SCHOOLS ¹

Score <i>m</i>	Frequency <i>f</i>	Score×Frequency <i>fm</i>
4	3	12
5	4	20
6	21	126
7	55	385
8	85	680
9	196	1,764
10	46	460
11	102	1,122
12	44	528
13	39	507
14	11	154
15	4	60
16	4	64
		614)5,882
		9.58 mean

It should be noted that the true mean is found in this case the same as in Table XVII. The principles underlying the computation of the simple and weighted means are the same. In each case the value of each measure, or score, is multiplied by its frequency; the products are added; and their sum is divided by the number of measures.

The formula for the weighted arithmetic mean is

$$M = \frac{\Sigma fm}{N}$$

where *M* equals the arithmetic mean, *m* the numerical value of any measure, *f* the corresponding frequency of occurrence, Σ the sum of *fm*'s and *N* the total number of measures.

¹ E. P. Cubberley, *School Organization and Administration*, p. 154.

In Tables XVII and XVIII, the data are said to be ungrouped and the exact value of each score is recorded; that is, there were three pupils who made a score of 4 (Table XVIII), four who made a score of 5, and so on. It so happens that the range on the Thorndike Handwriting Scale is from Quality 4 to Quality 18 inclusive, making a total range of 15. If, however, the range were extended, or the samples graded on a percentage basis, the grades might range from 0 per cent to 100 per cent. In that case the distribution might have 100 different scores instead of 13, as shown in Table XVIII.

Making a distribution table with from 80 to 100 different scores in it would involve a great amount of labor. In order to facilitate matters and lessen the volume of labor, the data are grouped into class intervals. The number of class intervals should rarely exceed 20 and a less number is, many times, desirable. The only reason for grouping the data into class intervals is to lessen the labor in computing the arithmetic mean. By this method, however, accuracy is sacrificed somewhat to save labor; but the difference between the true arithmetic mean with the data ungrouped and the mean found by this method (that is, with the data grouped) is so small that it is generally negligible.

In computing the arithmetic mean two assumptions must be made when the data are grouped in class intervals: (1) that the measures are distributed uniformly throughout the class interval; and (2) that for purposes of computation the measures in any class interval may be numerically represented by the *mid-point* of the class interval.

Two methods may be employed in the solution of the arithmetic mean with grouped data. Table XIX represents the data in Table XVIII grouped in class intervals and computed by the traditional, or *long method*.

TABLE XIX.—DISTRIBUTION OF SCORES ON 614 SAMPLES OF PENMANSHIP MADE BY CHILDREN IN THE THIRD GRADE IN THE SALT LAKE CITY PUBLIC SCHOOLS TO ILLUSTRATE THE COMPUTATION OF THE MEAN WITH DATA GROUPED IN CLASS INTERVALS

Class Intervals	Mid-point of Class Intervals m	Frequency f	Measures \times Their Corresponding Frequency fm
16-17.99	17	4	68
14-15.99	15	15	225
12-13.99	13	83	1,079
10-11.99	11	148	1,628
8- 9.99	9	281	2,529
6- 7.99	7	76	532
4- 5.99	5	7	35
		614	614 $\overline{)6,096}$
			9.92 <i>weighted arithmetic mean</i>

The *true arithmetic mean*, as shown in Table XVIII, is 9.58, while the arithmetic mean computed with grouped data is 9.92, making a difference of 0.34 of a score.

Computation of Arithmetic Mean by Short Method.—Table XX illustrates the computation of the arithmetic mean by the short method; the data are taken from Table XVIII.

The usual method in arranging the frequency distributions is to begin with the highest scores and work in the direction of the lower ones, that is, in Table XX, we place the scores from 16 to 17.99 at the top of the distribution. This is not absolutely necessary, but it is more convenient and less liable to errors.

When the data are grouped in the class intervals, as illustrated in Table XX, we may then take the mid-point of any class interval as the *assumed mean*. It is best to take the class interval that contains the true mean although this is not necessary. The point 9, the mid-point of the interval 8 to 9.99, was chosen as the assumed mean. The

TABLE XX

Class Interval	Frequency <i>f</i>	Deviation from the Assumed Mean Interval <i>d</i>	Frequency \times Deviation <i>fd</i>
16-17.99	4	+4	16
14-15.99	15	+3	45
12-13.99	83	+2	166
10-11.99	148	+1	148
8- 9.99	281	0	
6- 7.99	76	-1	-76
4- 5.99	7	-2	-14
	614		<u>+285</u>

$$285 \div 614 = 0.46$$

$$0.46 \times 2 = 0.92 = c, \text{ the correction}$$

$$\text{Assumed mean} \dots 9.0$$

$$\text{Correction} \dots \dots 0.92$$

$$\text{True mean} \dots \dots 9.92$$

mid-point of the class interval immediately above is 11 and is said to have a deviation (*d*) from the assumed mean of +1. The second class interval above has a deviation of +2 and so on. The mid-point of the class interval immediately below the assumed mean has a deviation of -1; the second one a deviation of -2, and so on. We next multiply the *deviations* (*d*) by the *frequency* (*f*) just as we did in the long method. This gives us the *fd* column in the table. It is evident that, if the assumed mean occupied the same position as the true mean, the sum of the deviations above it would equal the sum of the deviation below, but since the true and the assumed means are not the same, the measures above will not equal those below. Of course, it is possible that the assumed mean might equal the true mean, in which case there would be no correction; but this would rarely happen. In most cases, therefore, a *correction* (*c*) must be added to the assumed mean to get the true mean. This correction is

the *algebraic* sum of the fd 's divided by the total number of cases in the distribution. It is expressed by the formula

$$c = \frac{\Sigma fd}{N}$$

in which c equals the correction to be added, Σfd equals the algebraic sum of the frequencies (f) multiplied by their respective deviations (d), and N is the number of cases in the entire distribution.

The *average amount* of deviations from the assumed mean, taken in the right direction (that is, added algebraically) evidently would give us the *true mean*. It is evident that, if the sum of the plus fd 's is greater than the sum of the minus fd 's, the true mean deviates from the assumed mean in the direction of the positive fd 's, or the true mean is greater than the assumed mean by an amount equal to the correction c . If, however, the sum of the negative fd 's is greater than the sum of the positive fd 's, then the correction must be subtracted, and the true mean is less than the assumed mean.

In Table XX the positive fd 's exceed the negative fd 's by 285. This divided by the total number of cases (614) gives 0.46 of a class interval that each of the measures is in error, when the mean is considered to be at 9, the mid-point of the class interval. Since the width of the class interval is 2, we multiply the average error 0.46 by 2, giving 0.92 of an actual unit that must be added to the assumed mean. Thus we have: $9 + 0.92 = 9.92$ the true mean.

It should be noted that the short method is short only when the range is long and, therefore, many class intervals are necessary. If the range is short, there is nothing gained by using the short method. Following is the summary of steps used in this method:

SUMMARY OF STEPS IN THE COMPUTATION OF THE ARITHMETIC
MEAN BY THE SHORT METHOD

1. Group the measures in a frequency distribution table. Arrange the data in the table in four columns. The first column at the left contains the *class intervals* arranged with the highest scores at the top of the table; the second column contains the *frequencies* (f); the third column contains the *deviations from the assumed mean* (d); the fourth column contains the products of the frequencies times the deviations (fd 's).

2. Find the total of the frequencies in the second column.

3. By inspection estimate the class interval that contains the mean and take as the *assumed mean* the mid-point of this interval. (The mid-point of any class interval may be taken as the *assumed mean*; but it is better to choose the class interval containing the true mean or one near it.)

4. Consider each class interval as a unit and record in the third column the number of units that the mid-point of each class interval deviates from the assumed mean; the first one above the assumed mean having a deviation of $+1$; the second one a deviation of $+2$, and so on. Deviations below the assumed mean are treated in the same way, but considered negatively.

5. Multiply each deviation (d) by its corresponding frequency (f) observing the algebraic signs, and record the product in column 4.

6. Find the algebraic sum of the fd 's in column 4.

7. Divide this sum (which is the difference between the positive and negative fd 's) by the total number of measures (N) which is the sum of the frequencies in column 2. This gives the arithmetic mean of the deviations from the assumed mean in terms of class intervals.

8. Multiply the deviation in terms of class intervals by the number of units in the class interval.

9. Add (algebraically) the product obtained in 8 to the assumed mean to get the true mean.

The advantages and disadvantages of the different forms of average will be taken up after the *mode* and *median* have been discussed.

The Mode.—The mode is that item or term that is most characteristic or frequent in a distribution. It is the value

which is the fashion (*la mode*). It represents the typical fact. In Table XVII, page 280, there is no mode because one item or grade occurs as often as another. In Table XVIII, score 9 represents the mode because that score occurs more frequently than any other.

A mode may be defined as that measure of a variable fact which appears more frequently than measures directly above or below it. Distributions may, therefore, be unimodal or multimodal. A symmetrical distribution is unimodal because there is only one place in the distribution where the measures are of greater frequency than those directly above or below it.

The value of the mode as a measure of central tendency over the average may be illustrated by the following example: Suppose one were told that the average wealth of ten farmers living on a certain highway was over \$100,000 each. It might so happen that one of these farmers was worth \$1,000,000 and the other nine were worth \$1,000 each. The average or mean used in this case is misleading. The mode would be a much better average to use. In this case the mode would be \$1,000, which represents the group better than the mean which would be more than \$100,000. The mode has little statistical value other than an inspectional average and will, therefore, be discussed very briefly.

The Median.—The fact that the median has not been rigorously defined, or, if thus defined, the definition has not been generally accepted, has led to considerable confusion in its computation.

Rugg says that the median is "defined rigorously as that point on the scale of the frequency distribution on each side of which one half of the measures fall."² The measures, of course, must be arranged according to their

² Harold O. Rugg, *Statistical Methods Applied to Education*, p. 104.

ascending or descending values. While Rugg calls this a rigorous definition, it, nevertheless, has admitted of many solutions because of different interpretations of the limits of class intervals and the lack of uniformity in the distribution of the cases over them. While theoretically there would be no class intervals near the center of the distribution that contains no measures, yet actually this happens many times in practice when the number of measures in the distribution is small.

In practice we have not been rigorously consistent in the scoring of test papers and the computation of the median from the scores thus obtained. Teachers and students have had considerable trouble in finding medians, because in actual practice a half dozen methods are being employed, no two of which will give the same result. In order to help clarify the procedure and point out the basic principles and assumptions made, we shall discuss the computation of the median somewhat in detail. Before doing this, however, we shall give some additional definitions of the median in order to have it clearly in mind as the discussion proceeds.

Thorndike defines the median thus:³ "The median, or 50 percentile or mid-measure is the place on the scale reached by counting half the measures, in the order of their magnitude, or the place on the scale above and below which are equal numbers of the measures." This definition and the one given by Rugg make no provision for the computation of the median when the number of measures in the distribution is even and there is no middle measure. As a consequence, statisticians have used a variety of methods in computing medians with distributions of this kind.

Secrist defines the median thus:⁴ "The median of a

³ Edward L. Thorndike, *Mental and Social Measurements*, pp. 36-37.

⁴ Horace Secrist, *An Introduction to Statistical Methods*, p. 238.

series is that item—actual or estimated—in a series, when arranged consecutively, which divides the distribution into equal parts. When the number of items is even, it is half way between the two middle terms; when the number is odd, it is the middle term.” This definition is the same as that used by McCall who defines the median thus:⁵ “When measures are arranged in order of size, the median is the middle measure or (lacking a middle measure) midway between the two middlemost measures.” A median thus defined is perfectly definite and admits of but one interpretation provided there is uniformity in the treatment of the measures in the class intervals. We shall now note wherein the methods have differed in the computation of the median.

1. *The Spread of the Score Interval Commonly Used in Statistics and School Practice.*—When a pupil takes a test in addition in arithmetic, for instance, what are the limits of the various scores? That is, does a score of 2 mean that the pupil has done a definite amount of work, just barely finished two problems, and no more, or does it mean something else? What do five problems mean? The custom is to give no credit unless the pupil completes at least one problem. If he does less than one problem, he is given a grade of zero. Therefore, zero means, in actual practice, anything between just not any of the thing in question and 1. One problem means any amount between 1 and 2, but not exactly 1 and no more. Five problems means any amount between 5 and 6, and so on. We thus use the lower limit of the step, or score interval, in recording grades in arithmetic and in most of the other subjects in the curriculum. In finding the median score in most of the school achievement tests, however, the custom has been to use the middle of the step in computing the median and

⁵ William A. McCall, “How to Compute the Median,” *Teachers College Record*, Vol. 21, March, 1920, p. 126.

the lower limit of the score interval in scoring the papers and recording the scores.

From the standpoint of statistics, the middle of the score interval is the best measure to use. But the score interval generally used in statistics is not the same as the one used in actual practice in scoring test papers. From 1 to 2, 2 to 3, etc., are the score intervals used in scoring papers. But in the computation of the median the common practice is to employ score intervals with their limits at 0.5, 1.5, 2.5, etc. Statistical treatment is not consistent and does not conform to the practice in at least three respects: (1) Statistical methods use the mid-point of the score interval in computing the median but use the lower limit of the score interval in scoring the papers. (2) In statistics one score interval is used in computing the median, and another is used in scoring the papers and recording the grades. (3) In statistics the practice has been to use a different score interval in computing medians from that used by teachers in scoring test papers and recording the grades. In the following pages we shall note how statistical methods may be made more consistent and at the same time conform to the methods used by teachers in scoring papers.

2. *What Formula Shall We Use in Computing a Median?*—Some authors recommend the use of the formula $\frac{N+1}{2}$ in finding the median, and others recommend the

formula $\frac{N}{2}$. Both formulas will not give the same result in all cases. The question then arises as to which formula should be used, and if both are to be employed, when shall the former be used, and when shall we use the latter?

It seems to the author that there is no very good reason for using the formula $\frac{N+1}{2}$ since the same point on the

scale is not reached in computing the median if we employ this formula and count in from each end of the series. Furthermore, if we make the score intervals consistent with our logic in scoring papers, then the formula $\frac{N+1}{2}$

cannot be employed at all. It is sometimes argued that if you want the *ordinal* number of the median you should use the formula $\frac{N+1}{2}$ but if you want the *median point*

on the scale, the formula $\frac{N}{2}$ should be used. In answer to this argument it may be shown that both the ordinal number and the median point on the scale may be found by the latter formula. We shall now illustrate the various cases in the computation of the median.

Computation of the Median, Simple Distribution.—

CASE 1. *The number of items in the distribution is odd.*

TABLE XXI.—SCORES MADE BY THIRTEEN SIXTH-GRADE PUPILS IN THE COURTIS ARITHMETIC TESTS, SERIES B

Pupil	C	M	L	A	D	B	F	J	H	I	K	E	G
Score	16	15	14	13	12	11	10	9	8	7	6	5	4

Since the median is the measure in the middle score interval, it may be located by dividing the score intervals by 2. (The score interval here is the same as the class interval since there is only one score in a class interval.)

$\frac{N}{2} = \frac{13}{2} = 6.5$. In our reasoning let us make the score

intervals conform to practice in scoring papers and assume that they are from 4 to 5; 5 to 6, etc. Starting with the score interval 4-5 and counting 6.5 score intervals locates the median at 10.5. Counting down from the other end of the series we reach the same point. This is both the

mid-point of the middle measure (a measure is assumed to be distributed over a score interval) and the mid-point on the scale. The middle score is the score made by pupil F. Its mid-point is halfway between 10 and 11 or 10.5. Therefore 10.5 is the median. The customary way of computing the median is to take the score intervals one-half interval lower than these and make 10 the median score instead of 10.5. This, however, is inconsistent with the method of scoring the papers.

CASE 2. *The number of items or scores is even.*—Let us eliminate pupil G with a score of 4 from the preceding distribution and compute the median with an even number of cases. Table XXI now becomes Table XXII.

TABLE XXII.—SCORES MADE BY TWELVE SIXTH-GRADE PUPILS IN THE COURTIS ARITHMETIC TESTS, SERIES B

Pupil.....	C	M	L	A	D	B	F	J	H	I	K	E
Score.....	16	15	14	13	12	11	10	9	8	7	6	5

It is evident that the median score cannot be a middle score in this series since there is no middle score. Applying the formula used in Case 1 we have $\frac{N}{2} = \frac{12}{2} = 6$. Since

there is no middle score, some provision must be made to satisfy a case of this kind. Since there are 12 score intervals, it is evident that if we took the junction point of the 6th and 7th score interval we would have the mid-point on the scale. Starting with the 5-6 interval and counting in six intervals we have 11 as the mid-point on the scale. This value is obtained by counting in from either end of the distribution. The usual method of representing a score interval in statistics is by its mid-point. The mid-point of the 10-11 score interval is 10.5 and the mid-point of the 11-12 score interval is 11.5. If we take

half of these two middlemost values, we get 11 as the median score which conforms to the definition given by McCall and Secrist. This method is logical and conforms to the way scores are computed from test papers. If, as has been the custom, the score intervals are taken at 6.5 to 7.5; 7.5 to 8.5; etc., then the median score would be 10.5 instead of 11. This method, however, is inconsistent, as was pointed out above.

When the Distribution Is Complex.—CASE 3. *Where more than one pupil make the same score and the data are grouped in class intervals.*—The distribution of educational data is rarely so simple as those given in Tables XXI and XXII. Instead of a score being made by just one pupil it is usually made by more than one, sometimes by hundreds of pupils. This necessitates grouping the data in *class intervals* in order to condense the distribution table within workable limits. It also involves the distribution of the items within the class intervals.

The question of discrete and continuous series discussed in Chapter IX should be reviewed in order to have clearly in mind the type of data under consideration. Most measurements in education belong to a continuous series or may be treated as continuous even though discrete.

Having decided the question as to whether the data are discrete or continuous, the next important question to decide is the distribution of the items in the various class intervals. Suppose we had a group of 50 children who made a score of 7 in arithmetic on the Courtis arithmetic test. The chances are that some of the pupils had the eighth problem almost finished, when the signal to stop was given. Others had it three-fourths finished, some had it half finished, some one-fourth finished, and a few had just barely finished the seventh problem when time was called. In other words, instead of the 50 pupils just barely finishing the seventh problem the instant that time was

called, the probabilities are that they were distributed about equally over the interval 7 to 7.9999+. Instead of having 50 pupils make a score of 7, let us simplify the problem and take a distribution where four pupils make a score of 7 and the median falls within the 7-8 interval.

TABLE XXIII

Pupil.....	C	M	L	A	D	B	F	J	H	I	K	E	G
Score.....	9	9	9	8	8	7	7	7	7	6	5	5	4

The number of scores is 13. The median score is the seventh score counting in from either end of the series. The score made by pupil F is therefore the median score. But there are four pupils who made a score of 7. The best guess we can make as to the way those four scores are distributed in the class interval 7-7.9999+ is to say that they are distributed equally over the class interval.

The following illustration will make it clear what is meant by being distributed equally over the class interval. The table shows that pupil G had solved four problems; pupil E, five; pupil K, five; pupil I, six; pupils H, J, F, and B, seven, and so on, when the examiner gave the signal to stop. It does not mean, however, that just the instant the signal to stop was given that pupil G had just barely finished four problems and had not started on the fifth one, or that pupils E and K had just barely finished five problems and had done no work on the sixth, and so on. These scores indicate the number of problems actually completed and each of the 13 pupils may or may not have attempted more problems than those actually completed. Since the interval 7 to 7.9999+ is the interval that contains the median, the problem is to find out what is the most probable amount of work done by the seventh pupil counting in from either end of the series. Let us represent

graphically the 7-8 interval. We know that four pupils completed seven problems; but we do not know how much more they did. Now, our best guess would be that the actual amount of work done by these four pupils is distributed somewhat as follows:

Let us represent the class interval 7.00 to 7.9999 \pm .8 by Figure VII and let the line *AB* represent the distance through this step. When the signal to stop was given,

the first of these four students (student

H) had completed 7 problems and had

done work on the eighth one ranging in

amount somewhere between 0 per cent

and 25 per cent; J, the second student,

had the eighth problem from 25 per

cent to 50 per cent completed; F, the

third pupil, had completed from 50 per

cent to 75 per cent of the eighth prob-

lem, and B, the fourth pupil, had com-

pleted from 75 per cent to 100 per

cent. This would be a safer guess than

to guess that all four of these students had just barely

completed the seven problems when the examiner gave the

signal to stop. Still another guess is necessary. Assum-

ing that we know that the first pupil in the 7-8 interval

had done work on the eighth problem ranging in amount

somewhere between 0 per cent and 25 per cent, and that

we wanted to make the best estimate we could make, tak-

ing one case with another, as to how much he actually

did, we would estimate that he had gone halfway through

this interval 0 per cent to 25 per cent, or, that he had

done 12.5 per cent of the eighth problem when the signal

to stop was given. Reasoning the same way for the other

score intervals, the second pupil (J) would have had 37.5

per cent of the eighth problem completed; the third (F)

62.5 per cent, and the fourth one (B) 87.5 per cent.

FIGURE VII

B	7.9999 - .8
B	-87.5
E	-7.75
F	-62.5
D	-7.50
J	-37.5
C	-7.25
H	-12.5
A	7.00

Now, since we have the most reasonable distribution of these cases over the class interval, we are ready to continue the work in computing the median. Starting at the lower end of the distribution we have four cases up to the 7-8 interval. Since the median score is the score made by the seventh pupil, and since there are 13 score intervals in the entire series, therefore, counting in from either end of the series $\frac{N}{2}$ score interval, we would have half the distance through the series and also have the mid-point of the seventh score interval. We have four cases up to the 7-8 class interval and must have 2.5 of the 4 score intervals in the 7-8 class interval. Noting Figure VII, we see that 2.5 score intervals would give us a point 62.5 per cent through the 7-8 class interval, or a median point on the scale of 7.625. But 7.625 is also the mid-point of the seventh score interval. Therefore the median is 7.625.

Table XXIV illustrates the computation of the median with the data grouped in class intervals, each of which contains five units instead of one, and with some of the class intervals containing a large number of cases.

TABLE XXIV.—DISTRIBUTION OF MARKS GIVEN IN ENGLISH TO 263 HIGH-SCHOOL PUPILS

Class Interval	Number of Pupils
95.0-100.00.....	20
90.0- 94.99.....	63 (83) Adding down
85.0- 89.99.....	38 (121)
80.0- 84.99.....	47
75.0- 79.99.....	38 (95)
70.0- 74.99.....	33 (57)
65.0- 69.99.....	16 (24)
60.0- 64.99.....	2 (8)
55.0- 59.99.....	3 (6)
50.0- 54.99.....	1 (3)
45.0- 49.99.....	1 (2) Adding up
40.0- 44.99.....	1

N = 263

Applying formula, $\frac{N}{2} = \frac{263}{2} = 131.5$.

Since there are 263 cases, 131.5 score intervals from either end of the series will be the mid-point on the scale and also the mid-point of the middlemost measure.

Adding up from the bottom, we have 95 cases up to the 80-84.99 class interval. This class interval contains 47 cases, or 47 score intervals, and we must take 36.5 of these score intervals to give us the median point on the scale, and the mid-point of the middlemost measure. The same result is obtained by counting down from the top. There are 121 scores down to the upper margin of the class interval 80.0-84.99. Therefore, we must take 10.5 scores from the 47 in that interval or go $\frac{10.5}{47}$ the distance through the class interval coming in from the top. Since the number of units in the interval is five, $\frac{10.5}{47}$ of 5 subtracted from 85, the upper limit of the interval, equals 83.88, which is the median.

CASE 4. *Where the median falls in the 100 or the zero class interval.*

TABLE XXV.—DISTRIBUTION OF THE MARKS GIVEN TO A FRESHMAN CLASS IN ALGEBRA

Class Interval	Frequency
100.....	13
90-99.99.....	1
80-89.99.....	2
70-79.99.....	1
60-69.99.....	1
50-59.99.....	2
40-49.99.....	1
30-39.99.....	1
20-29.99.....	1
	<hr/> 23

Substituting in the formula, $\frac{N}{2} = \frac{23}{2} = 11.5$. Counting from either end of the series we find that the median lies in the 100 interval. The cases in this interval are undi-

tributed and are considered to lie at a point; therefore there is no correction, and the median is 100.

In the solution of problems no credit is usually given unless the student solves one or more problems correctly. In a class of 25, it might be that 13 pupils would fail to complete any problems. In this case the median would depend upon whether the cases in the zero-interval, that is, from 0 to 0.999 +, were distributed equally over the interval or whether they were considered to be piled up at its lower limit. From the reasoning in the former cases it would be more logical to consider them distributed over the interval and take $\frac{13}{12.5}$ of the distance through the interval as the median measure. Therefore, the median is 0.96.

CASE 5. *When the partial sum is the half sum and there is no correction.*—Another case that proves troublesome is illustrated in Table XXVI, taken from Monroe.⁶

TABLE XXVI

Scale	Frequency
15.....	
14.....	
13.....	1
12.....	1
11.....	2
10.....	3
9.....	5
8.....	4
7.....	6
6.....	3
5.....	
4.....	
Total	25
Approximate median.....	9.0
Correction.....	0.0
True median.....	9.0

⁶ *Measuring the Results of Teaching*, p. 106.

In explaining the median in this case, Monroe says: "Case A is where the partial sum (13) is also the half sum. The approximate median is in the *next* interval (9). Since the difference between the partial sum and the half sum is zero, there is no correction and the true median is 9.0."

This practice does not conform to the theory discussed above. Neither is it consistent with the instructions that Monroe gives for finding the median in his reading tests.

Reasoning as we did from Table XXIII, we note that since the interval 8-9 contains four measures which are theoretically distributed equally over the interval, the thirteenth measure would lie in the class interval somewhere between 8.75 and 8.9999 and our best guess is that it lies at the mid-point of that interval or at 0.875 of the distance through the class interval, which would make the median 8.875 instead of 9.

CASE 6. *Where measures are discrete.*—Table XXVII illustrates the computation of the median with measures discrete.

TABLE XXVII.—ILLUSTRATING THE COMPUTATION OF THE MEDIAN, SERIES DISCRETE

Number of Pupils in Class	Frequency
10.....	14
11.....	17
12.....	21
13.....	27
14.....	20
15.....	18
Total.....	117

Since there are 117 measures the 59th measure is the median measure. Counting up from the bottom we find that the 59th class lies somewhere among the classes containing 13 pupils. Now since it is, of course, impossible

for a class to contain a fraction of a pupil, the median is 13.

CASE 7. *Where the median falls within a class interval containing no cases.*

TABLE XXVIII

Scale	Frequency
11.....	8
10.....	10
9.....	20
8.....	0
7.....	10
6.....	13
5.....	10
4.....	5
Total.....	<hr/> 76

$\frac{N}{2}$ equals 38. There being an equal number of cases, the median is mid-point between the two middle cases. But there is one class interval between the two middlemost cases that contains no measures. Since the number of cases is even, the median point on the scale would ordinarily be the junction point of the two middlemost score intervals, but, since the two middlemost score intervals are not contiguous, we add half the intervening class interval to the score interval above and the other half to the score interval below and locate the median at the mid-point of the gap between the two middlemost score intervals. Halfway through the 8-9 interval, therefore, or 8.5, is the median.

A somewhat lengthy discussion has been given on the computation of the median, because the median is one of the chief measures of central tendency, and also because the methods are not uniform. We shall now give a summary statement of the steps for the computation of the median.

SUMMARY OF STEPS IN THE COMPUTATION OF THE MEDIAN

1. Arrange the data in a frequency distribution taking special care to note the limits of the class intervals and also the limits of score intervals.

2. Find one-half the sum of the measures.

3. Beginning at either end of the distribution, preferably the lower end, count the number of measures included in all class intervals up to the interval containing the median.

4. Subtract this number from the half sum of all the measures computed in step 2. The difference is the number of measures that must be taken from the next interval to bring the computation up to the median point on the scale.

5. Divide this remainder by the number of cases in the class interval containing the median and multiply the quotient by the number of units in the class interval.

6. Add this number to the value of the lower limit of the class interval containing the median, if computation is made from the lower end of the distribution, and subtract it from the upper limit of this class interval, if computation was made from the upper end of the distribution. This is the median point on the scale.

If, in finding the median, no other measures are desired, care need be taken only in the arrangement of the cases near the median value. Consequently the median cannot give detailed information of the measures at the extremities of the ranges. On the other hand, the median is a fairly stable measure and changes very slowly when different samples are taken, which means that it is not greatly affected by the presence of accidental and irrelevant influences.

Comparison of the Arithmetic Mean, Mode, and Median.

—We are now in a position to compare the three kinds of averages with a view of determining their more salient characteristics. Following is the comparison of the arithmetic mean, the mode, and the median:

ARITHMETIC MEAN

1. The arithmetic mean takes into consideration all cases and is affected by their size.
2. The arithmetic mean is affected by every item in the group.
3. The mean may be found without arranging the measures according to their magnitude.
4. The mean may be determined when the aggregate and the number of cases are known.
5. The mean may fall where no data actually exist.

MODE

1. The mode deals with only the most representative measures and neglects the extreme cases.
2. The mode is determined by the most frequent measures only.
3. The measures must be arranged according to their magnitude.
4. The mode may be located without the number of cases or the extreme cases.
5. The mode falls where the cases are most numerous.

MEDIAN

1. The median takes into consideration all the cases; but the size of extreme cases does not affect it.
2. The median is a counting average and is affected by the number of cases, but not by the size of the extreme cases.
3. The measures must be arranged according to their magnitude.
4. The sum of the measures and their number do not furnish sufficient data to compute the median.
5. The median, like the mean, may be interpolated, and fall where no case actually exists.

A comparison of the above averages indicates that the nature of the data and the problem to be solved must determine the average to be used. If the size of the measures and the number of cases are to be taken into consideration, then the arithmetic mean is the average to use. If, however, the most characteristic measure of the group is wanted, the mode best satisfies this condition. The arithmetic mean has the advantage of being a common measure and one with which the public is familiar. Its calculation is simple, but it is greatly effected by extreme cases and for that reason it should many times give way to the median or mode. One disadvantage of the mode is the fact that there is many times no well-defined type, and one

measure appears as often as another. In this case the median is probably the most representative term.

Quartiles and Percentiles.—It is sometimes convenient to divide the distribution into divisions smaller than those made by the median, that is, to divide it into quarters, tenths, etc. The medians dividing the halves of the distribution into equal parts are known as *quartiles*. Starting with the lower end of the distribution, the median dividing the lower half is known as the *first quartile*, (Q_1), whereas the median dividing the upper half of the distribution is known as the *third quartile*, (Q_3). The computation of the quartiles is the same as the median except that in the first quartile we take the $\frac{N}{4}$ case and in the third quartile the $\frac{3N}{4}$ case instead of the $\frac{N}{2}$ case, as in the median.

In the same way we may find any desired *percentile* in the distribution. If the distribution is divided into ten equal parts, the division points are known as *deciles*. A series of such measures gives a more complete picture of the distribution than can be obtained from a single measure.

CHAPTER XI

MEASUREMENTS OF DISPERSION, OR VARIABILITY

In the previous chapter we noted the measures of central tendencies. We shall now note the dispersion or variation from these measures. Measures of variation call special attention to the degree of homogeneity which characterizes the distribution. The simplest measure of dispersion is the range, that is, the difference between the greatest and least magnitude in the series. While this is a simple measure of dispersion, it is a very imperfect one, as will be shown later, because two distributions might have the same range and yet differ widely in their "scatteration." A few illustrations will make clear the point why measures of variability are necessary in describing a group of data.

Suppose a teacher who contemplated going to another state to teach school was told that the mean salary of teachers in that state was \$1,250. It is evident that the mean does not convey sufficient information to make one intelligent as to what his chances are of getting the mean salary. It might be that nine-tenths of the teachers received salaries between \$1,200 and \$1,600, in which case one's chance of getting a salary pretty close to the average would be good. Or it might be that no teacher got a salary within \$200 of the average. The measure of central tendency does not, therefore, give one anything like exact information as to what to expect.

When one says that the average grade or the median grade of a freshman class in algebra is 75 per cent, it does not indicate the distribution of the grades in that class. They

may range from 5 per cent to 100 per cent, or from 65 per cent to 80 per cent, or cover any other range from 0 per cent to 100 per cent. The description of the distribution is far more nearly complete when the dispersion or "scatteration" is given.

At best the average is but a partial measure of type. If one desired to compare the salaries of teachers in two states or the grades in two schools and had nothing but the measures of central tendency, it would be impossible to draw any definite conclusions, because the forms of the distributions might vary greatly. Two distributions might have the same range and the same mean, median, and mode and still differ widely as to form. In one distribution the cases might be concentrated near the measure of central tendency while in the other they might be distributed about equally over the entire range. Or, again, the measures in one distribution might be concentrated at the ends of the range while in another distribution they might be concentrated at or near the center.

How Variability Is Measured.—We noted in the previous chapter that a measure of central tendency was a *position*, or *point on the scale*. Variability differs from a measure of central tendency in that the latter is a *point* or a position on the scale whereas the former is a *distance*. Variability is expressed as the distance on the scale that will include a certain proportion of the measures in the distribution. This distance is expressed in various units, depending in part on the nature of the distribution and in part on the arbitrary choice of the statistician.

Measures of Absolute Variability.—There are four measures of absolute variability in common use. They are:

1. The *range*, which includes all of the measures in the distribution.
2. The *mean deviation* is the mean of all the deviations from a measure of central tendency, such as the median or

mean. When laid off on each side of the average in a normal distribution, it includes the middle half of the cases.

3. The *standard deviation* is the square root of the mean of the squares of all deviations when the deviations are measured from a measure of central tendency, either the mean or the median. When thus laid off, it includes approximately the middle two-thirds of the distribution.

4. The *quartile deviation, median deviation*. The quartile or median deviation applies to that portion of the distribution contained between the first and third quartiles and is computed by taking one-half the range contained in the middle half of the distribution. It may be computed from the formula

$$\frac{Q_3 - Q_1}{2},$$

where Q_3 represents the third quartile and Q_1 the first quartile.

Another term frequently used is the *probable error* discussed in a previous chapter. If the distribution is symmetrical then the probable error equals the median deviation and includes the middle half of the cases in the distribution. If the distribution is not symmetrical, but skewed, it is questionable whether the term should be used. The term really belongs in sampling. Sampling is used under the following conditions: In statistics it is usually impossible, or at least not convenient, to obtain all the measures of any group of things under consideration, so we take samples and judge the entire group by the samples taken. When a farmer brings his wheat to town, the man at the elevator does not examine critically the entire load but takes a sample from the front end of the load, one from the rear, and perhaps one from the middle, and judges the entire load from the samples taken. In a city school system a superintendent may desire to know what score the fourth-grade children

are able to make on a certain test. It may be that he does not have time to test the fourth grade throughout the entire city. He therefore chooses a few schools at random and judges the entire fourth grade by the samples taken. But to be scientific he must know how reliable his samples are. That is, if he had taken samples from other schools, would they have differed radically from those taken? How different would the results have been if he had tested the fourth grade through the entire city? *Probable error* is a means of testing the reliability of samples. It is a quantity such that we would obtain values of greater and less magnitude with equal frequency if more cases or samples were taken. If the distribution is normal, it is evident that there will be as many measures greater than those in the middle half as there are those that are less. Or, it is a "fifty-fifty" chance, when measures are under consideration not included in the middle half, that they will be larger than those in the middle half as often as they are smaller. For example, suppose a city superintendent desired to know the mean score of the eighth-grade children on the Monroe Silent Reading Test and that he did not have time to test all the eighth-grade children throughout the entire city or that the expense were too great. He might then test ten classes, for instance, and determine the mean grade of the classes chosen. Let us suppose that the mean score was 25. He would next want to know how near this score of 25 would be to the mean score if he had tested all of the eighth-grade children throughout the entire city. He knows that as he went from school to school testing the eighth grades, the mean score of some of them was more than 25 and for others it was less. In making his calculations let us suppose that he found the probable error to be 6. This means that in taking his ten samples as often as he found a measure 19 or less (6 below 25), he would find one 31 or more (6 more than 25). It is evident that the greater the difference in

reading ability found in the samples taken, the greater the probable error. In other words a small P.E. means a rather homogeneous group. It is also evident that P.E. would not be the correct measure to use unless the distribution were normal. Since the quartile deviation is determined by counting the number of measures between the first and third quartiles, and taking one-half of them, it is really not deviation at all.

Computation of the Mean Deviation.—We have indicated above that the quartile deviation is not a deviation from any particular average and takes account of the form of the distribution only in an indirect manner. The mean deviation, on the other hand, is a real measure of deviation from a measure of central tendency.

The illustration in Table XXIX will make clear the significance of the mean deviation.

TABLE XXIX.—GRADES MADE BY TEN PUPILS IN THE FIRST-YEAR ALGEBRA

Pupils	Grade	Deviation from Mean
A	79	6.6
B	91	5.4
C	76	9.6
D	86	0.4
E	88	2.4
F	90	4.4
G	95	9.4
H	94	8.4
I	78	7.6
J	79	6.6
	10)856	10)60.8
	85.6 = mean	6.08 = mean deviation

We note that the mean grade for this group of ten pupils is 85.6. No pupil makes the mean grade. The grade made by A differs from the mean 6.6. That made by B differs by 5.4 and so on. The sum of all the deviations

from the mean, without reference to signs (that is, without reference to whether they are above the mean or below it), divided by the number of cases, which is 10, gives the average or mean deviation, 6.08.

The mean deviation may be computed from either the mean or the median. It would be the same from either if the distribution were symmetrical. If only slightly unsymmetrical, it would be the same to the second decimal point. If the distribution is considerably skewed, however, the deviation is less from the median because the arithmetic mean is affected by both the size of the items and the frequencies. In the case of the median only the frequencies and the size of the items near the center of the distribution affect this measure. The following illustration from Bowley will make clear the reason why deviations from the median are less than from the mean.¹

Suppose that it is required to run from a telephone exchange separate wires to every one of N places in a straight line, where should the exchange be placed so as to use the least total amount of wire? At the median position. For if you move from the median position to the right, or to the left, you will find immediately that you are adding more wire than you are subtracting. Supposing there are 20 stations, and you have a position between the 10th and 11th; if you move to a position between the 11th and 12th you have to increase your distance from ten stations and diminish it from nine, in every case by the same length of the wire. The wires correspond to the deviations; and the sum of the lengths of the wires is the sum of the lengths of the deviations.

It is evident that with an arithmetic mean there may be more cases on one side of the mean than on the other; therefore, the sum of the distances from the mean to the various cases would be greater than from the median. From a mathematical standpoint it would seem, therefore, that the median is the proper measure of central tendency to use in computing the mean deviation.

¹ A. L. Bowley, *Measurement of Groups and Series*, p. 30.

Computation of the Mean Deviation: Data Grouped in a Frequency Distribution.—In Table XXIX we illustrated the computation of the mean deviation where the number of cases was small and the data were ungrouped. We shall now illustrate the computation with data grouped in a frequency distribution. Table XXX illustrates the method.

TABLE XXX.—DISTRIBUTION OF SCORES GIVEN TO 288 HIGH-SCHOOL PUPILS IN PLANE GEOMETRY, ILLUSTRATING THE COMPUTATION OF THE MEAN DEVIATION BY THE LONG METHOD

Class Interval	Mid-point of Class Interval	Frequency <i>f</i>	Deviation <i>d</i>	Frequency × Deviation <i>fd</i>
95-100	97.5	20	14.91	298.20
90- 94.99	92.5	62	9.91	614.42
85- 89.99	87.5	49	4.91	240.59
80- 84.99	82.5	27	0.09	2.43
75- 79.99	77.5	48	5.09	244.32
70- 74.99	72.5	23	10.09	232.07
65- 69.99	67.5	18	15.09	271.62
60- 64.99	62.5	21	20.09	421.89
55- 59.99	57.5	9	25.09	225.81
50- 54.99	52.5	6	30.09	180.54
45- 49.99	47.5	3	35.09	105.27
40- 44.99	42.5	2	40.09	80.18
		<i>N</i> = 288		2,917.34

$$\frac{N}{2} = 144 \qquad \frac{2,917.34}{288} = 10.13 \text{ M.D.}$$

The true median is 82.59 (computation not shown here).

The computation of the mean deviation by the method given in Table XXX involves a great amount of work by reason of the fact that the exact deviations are taken from the true median. These, in most cases, are fractions that must be multiplied by the frequency. Much labor may be saved by assuming that the median is at the mid-point of the class interval in which the true median is located and by reckoning the deviation in terms of class intervals instead

of in terms of the units of class intervals. We may then make the proper correction for the assumptions made. That is, if the deviations are reckoned about an assumed median instead of the true median, the proper corrections must be made for the difference between the assumed and the true medians.

It will be found that the sum of the deviations about the assumed median is always less than those about the true median; hence *the correction must always be added.*

This may be illustrated by Figure VIII. Suppose we have a distribution the range of which is from 40 to 100 and that the range is divided into class intervals of five units each. Let us suppose that the true median is 82.59 as in Table XXX. The assumed median is at 82.5. Since the true median is 82.59, it means that there are as many pupils who receive grades above 82.59 as there are below it. But we make an assumption that the median is located at 82.5 and that all the measures in the interval 80-85 lie at the midpoint of this class interval. Therefore the 27 cases would

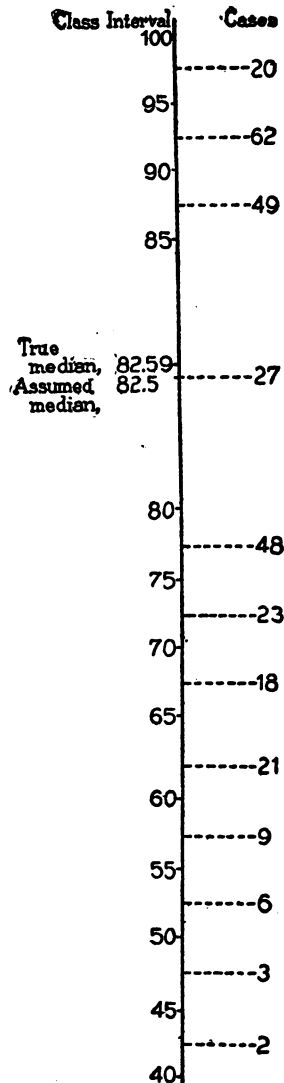


FIGURE VIII

be considered as lying below the true median and would be counted with those below. It is also evident that there are now more cases below the true median than above it, when the 27 cases in the 80-85 interval are assumed to be located at the mid-point, or at 82.5. This means that all the cases below the true median are too short by the difference between the true and assumed medians, or that they are short 0.09 of a unit or $0.09/5$ of a class interval, or 0.018 of a class interval. It also means that all the cases above the true median are too long by the difference between the

TABLE XXXI.—DISTRIBUTION OF SCORES GIVEN TO 288 HIGH-SCHOOL PUPILS IN PLANE GEOMETRY, ILLUSTRATING THE COMPUTATION OF THE MEAN DEVIATION BY THE SHORT METHOD

Class Interval	<i>f</i>	<i>d</i>	<i>fd</i>
95-100	20	3	60
90- 94.99	62	2	124
85- 89.99	49	1	49
80- 84.99	27	0	0
75- 79.99	48	1	48
70- 74.99	23	2	46
65- 69.99	18	3	54
60- 64.99	21	4	84
55- 59.99	9	5	45
50- 54.99	6	6	36
45- 49.99	3	7	21
40- 44.99	2	8	16
	<i>N</i> = 288		Σfd = 583

True median 82.59

Assumed median 82.5

131 number of cases above assumed median

157 number of cases below

26 difference

$$c = \frac{82.59 - 82.5}{5} = 0.018 \quad 26 \times .018 = 0.468$$

$$\frac{583 + 0.468}{288} = 2.026 \text{ in units of class intervals}$$

$$2.026 \times 5 = 10.13 M.D.$$

true and assumed medians. The number above is 131, since the cases in the 80-85 interval are counted with those below. The number below is 157. Therefore 131 cases are too long and 157 cases too short, and the difference between them, or 26 cases, are too short by 0.018 of a class interval, hence the correction must be added. If the true median had been below the assumed median, the number of cases above would have exceeded those below, the number of cases above would have been too short, and the correction would have had to be added as in the case cited.

Table XXXI illustrates the computation of the mean deviation by the short method. In this table we shall use the same data used in Table XXX.

If it is desired to use a general algebraic formula for finding the mean deviation by the short method, the operation may be expressed thus:

$$M.D. = \left[\frac{\Sigma fd + c(N_b - N_a)}{N} \right] \times 5$$

where $M.D.$ = the mean deviation, f = the number of cases in each class interval, d = the deviation in units of class intervals, c = the correction, which is the true median minus the assumed median, divided by the number of cases in the class interval; N_b = the number of measures below the true median, N_a = the number of measures above the true median, and N = the number of measures in the entire distribution. 5 = the number of units in the class interval. Substituting in the general formula to find the mean deviation in Table XXXI:

$N_b = 157$ True median = 82.59 Assumed median = 82.5

$N_a = 131$ $\Sigma fd = 583$ $c = \frac{82.59 - 82.5}{5} = 0.018$

$N = 288 \therefore M.D. = \frac{583 + 0.018(157 - 131)}{288} = \frac{583 + 0.018 \times 26}{288} = 2.026$

$2.026 \times 5 = 10.13 M.D.$

SUMMARY OF STEPS IN THE COMPUTATION OF THE MEAN
DEVIATION BY THE SHORT METHOD

1. Arrange the data in a frequency distribution of four columns. Let the first column to the left contain the class intervals; the second one the frequencies (f); the third the deviations (d); and the fourth the product of the frequencies times the deviations (fd).

2. Sum the frequencies in the f column.

3. Compute the true median (TM_d).

4. Take the mid-point of the class interval containing the true median as the assumed median (AM_d).

5. Find the difference between the true median and the assumed median and divide this difference by the number of units in the class interval. This will give the difference in terms of the class interval. Call this the correction (c).

6. Compute the number of cases above and below the true median in column f . Care must be taken to see whether the cases in the class interval containing the medians shall be added to the cases above the true median or below it. If the assumed median falls below the true median, then the case in the class interval containing the medians must be added to those below the true median for reasons given on page 311. If the assumed median falls above the true median, add the cases in the class intervals containing the medians to those above. Multiply the correction found in step 5 by the difference between the number of cases above and the number below the true median.

7. Tabulate the deviations (d) from the mid-point of the class interval containing the assumed median, giving the first class interval above a deviation of 1, the second one a deviation of 2, etc. Record the deviations below the interval containing the assumed median in the same way.

8. Multiply each frequency (f) by its respective deviation and record the results in the proper place in the fd column.

9. Find the sum of the fd 's without regard to sign.

10. Add the total correction from step 6 above to the total number of deviations from the median (Σfd).

11. Divide this sum by the total number of cases in the distribution (N) to get the mean deviation about the true median. The deviation thus computed is in terms of class intervals; therefore multiply this result by the number of units in the class interval in order to find the deviation in terms of the original measures.

The Computation of Standard Deviation.—We defined standard deviation (page 306) as the square root of the arithmetic mean of the squares of all the deviations measured from the arithmetic mean of the distribution.

If the series is simple and there is only one or a few measures of any given magnitude, the formula in this case is as follows:

$$\sigma = \sqrt{\frac{\sum d^2}{N}}$$

In substituting in this formula we simply find the amount each measure deviates from the mean, square it, add the squares of all the deviations, divide by the total number of cases, and extract the square root.

If, however, the number of cases is large, we arrange the data in a frequency distribution and apply the short method. The formula in this case is

$$= \sqrt{\frac{\sum fd^2}{N}}$$

in which σ = the standard deviation, Σ = the sum of the fd^2 's, f = the frequencies or the number of measures in each class interval, d = the deviations from the mean, and N = the number of measures in the whole distribution.

The Computation of Standard Deviation by the Short Method.—We shall note the short method of computing the standard deviation and shall then compare standard deviation with mean deviation, and note wherein one seems to be superior to the other. Table XXXII illustrates the computation of the standard deviation by the short method, and the steps in the procedure are given in the table on the following page.

TABLE XXXII.—DISTRIBUTION OF SCORES GIVEN TO 288 HIGH-SCHOOL PUPILS IN PLANE GEOMETRY, ILLUSTRATING THE COMPUTATION OF STANDARD DEVIATION BY THE SHORT METHOD

(Data taken from Table XXXI)

Class Interval	f	d	fd	fd^2
95-100	20	3	60	180
90- 94.99	62	2	124	248
85- 89.99	49	1	49	49
80- 84.99	27	0	233	
75- 79.99	48	-1	- 48	48
70- 74.99	23	-2	- 46	92
65- 69.99	18	-3	- 54	162
60- 64.99	21	-4	- 84	336
55- 59.99	9	-5	- 45	240
50- 54.99	6	-6	- 36	216
45- 49.99	3	-7	- 21	147
40- 44.99	2	-8	- 16	128
	$N=288$		-350	288)1,846
			233	6.41= S^2
			288)117	
			- .406	

$$c = -0.406$$

$$c^2 = 0.165$$

$$6.41 - 0.165 = 6.245 = \sigma^2$$

$$\sigma = 2.50 \text{ class intervals}$$

$$2.50 \times 5 = 12.50 \text{ actual units}$$

SUMMARY OF STEPS IN THE COMPUTATION OF STANDARD DEVIATION BY THE SHORT METHOD

1. Tabulate the data in a frequency distribution as in the computation of the mean deviation, adding a fifth column to the right to contain the fd^2 's.

2. Estimate the interval which contains the mean (80-84.99). This may be chosen anywhere in the distribution; but if chosen in the same interval, or near the true mean, the computation is simplified.

3. Tabulate the deviations from the estimated mean (the midpoint of the class interval) in units of class intervals; that is, the first interval above will have a deviation of 1, the second one a deviation of 2, and so on; the first interval below will have a deviation of -1, the second one will have a deviation of -2, and so on.

4. Multiply each frequency by its corresponding deviation and record in the fd column.

5. Find the algebraic sum of the fd 's; that is, find the sum of the $+fd$'s and the sum of the $-fd$'s, and take their difference: $\Sigma fd = 233 - 350 = -117$.

6. Find the correction (c) by dividing fd by the total number of cases in the distribution: $-117 \div 288 = -0.406$.

7. Multiply each fd by d , its corresponding deviation, and record in the column headed fd^2 . (The student should use the table of squares, Table I, Appendix, for squaring numbers, also for extracting roots.)

8. Find the sum of the fd^2 's: $\Sigma fd^2 = 1,846$.

9. Divide the sum of the fd^2 's by the number of cases in the whole distribution to get S^2 : $1,846 \div 288 = 6.41$. This is the square of the standard deviation, but it is computed from an estimated mean, and we must find it from the true mean. From the previous discussion it is clear that the mean of the deviations about the estimated mean must be in error by an amount equal to the arithmetic mean of the difference of the positive and negative deviations in column 4; that is, the arithmetic mean of the squares of the deviations will be in error by an amount equal to the squares of this difference, or, c^2 . Square the correction c , giving c^2 , or 0.165 and subtract c^2 from S^2 , giving σ^2 . It was noted in step 3 that the deviations were in terms of class intervals; therefore σ^2 will be in terms of class intervals, and its square root will, of course, be in the same terms. Multiplying σ by the number of units in the class interval (5) will give the desired standard deviation, 12.50.

The methods of computing the mean and standard deviations having been presented, we are now in a position to compare them and discuss the superiority of one over the other.

In the introductory chapter on statistical methods we devoted considerable space to the exposition of the arithmetical mean as the basis upon which rests the concept of error in a series of observations. We also discussed the method of *least squares* as a means of finding the most probable value of a series of observations. We are now prepared to show the significance of mean and standard

deviations graphically. This can be done best, perhaps, by using the illustrations given by Roberts.² Suppose A and B are two marksmen firing at a target the center of which is C, Figure IX.

Suppose each man fires ten shots. Let the crosses represent the hits made by A and the circles the hits made by B.

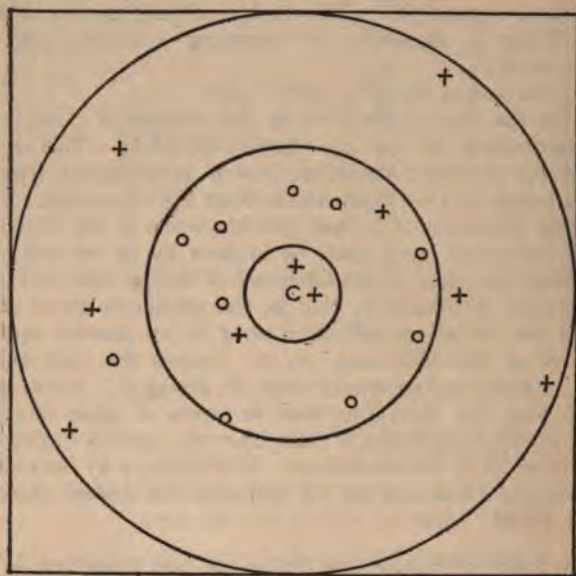


FIGURE IX. SHOTS FIRED AT A TARGET BY TWO MARKSMEN; ILLUSTRATING THE DIFFERENCE BETWEEN MEAN AND STANDARD DEVIATION

Table XXXIII shows the distance of each shot from the center of the target.

Considering the amount each man misses the target as a deviation, we find that the sum of the deviations of each marksman is 50 and the mean deviation is 5. Therefore,

² "A Practical Method for Demonstrating the Error of Mean Square," *School Science and Mathematics*, Vol. 19, pp. 677-692.

TABLE XXXIII

Shot Number	Distance of Shots from Center of Target, Inches	
	A	B
1	2	4
2	8	5
3	3	6
4	7	4
5	9	3
6	4	5
7	9	5
8	5	6
9	9	7
10	1	5
	<hr/> 50	<hr/> 50

if mean deviation were taken as a measure of their marksmanship, there would be a tie. If, however, we were to compute their marksmanship in terms of standard deviation, they would not tie, but B would be the winner.

Table XXXIV shows the standard deviation in each series.

Formerly the scores at rifle practice in the Belgian army were determined by adding the deviations of each man's shots from the center of the target, and the marksman having the smallest sum was the winner.

It is evident from looking at the scores made by A and B in Table XXXIV that B is a more consistent marksman than A because his marksmanship is less variable; that is, he "bunches his shots." At one time A fires a shot close to the target, and the next shot misses the target by many feet. B does not hit as close to the target as does A but is far more consistent in his shooting.

The mathematical significance of standard deviation may be further demonstrated as follows: In the United States army, the standard size target used on rifle ranges

TABLE XXXIV

Number of Shots	A		B	
	d	d^2	d	d^2
1	2	4	4	16
2	8	64	5	25
3	3	9	6	36
4	7	49	4	16
5	9	81	3	9
6	4	16	5	25
7	2	4	5	25
8	5	25	6	36
9	9	81	7	49
10	1	1	5	25
	<u>50</u>	<u>10)334</u>	<u>50</u>	<u>10)262</u>
		33.4		26.2

$\sqrt{33.4}$ = Standard deviation of A = 5.77

$\sqrt{26.2}$ = Standard deviation of B = 5.11

at 200-300 yards distance is rectangular in shape, 4×6 feet in dimensions, in the center of which are three concentric rings, of which the central one (called the bull's-eye) has a diameter of 8 inches, the second, a diameter of 26 inches, and the largest one, a diameter of 46 inches. In firing at a target of this kind the score is made up as follows:

First ring, or bull's-eye.....	5 points
Second ring.....	4 points
Third ring.....	3 points
Outside target.....	2 points

As indicated above, shots placed in a target may be taken in the sense of errors, or deviations from a mean. The mean in this case is the mathematical center of the target. The distance of each shot from the center is its deviation. This distance from the center of a target may be considered as a radius of a circle, and, since a shot may be placed at any point in a circle around the center of the target within the

limits of its radius, we may treat the various misses or deviations from the center as being proportional to the area of a circle πr^2 , of which r is the distance of the shot from the center. We may then determine the sum of the areas of the various circles formed by each man's shots taken as radii, and, from our reasoning above, the man having the smaller total sum of circle areas will be the winner. Or, stating the relative merits of their marksmanship in a mathematical formula we may say $A : B :: \Sigma(\pi r^2) : \Sigma(\pi r_1^2)$. π , being a constant, may be eliminated, and the formula may be stated thus $A : B :: (\Sigma r^2) : (\Sigma r_1^2)$. The rifleman having the smaller sum of squared deviations would win.

If, however, we wish to carry the measurements a step further and obtain a measure of the relative average fluctuation or miss from the center of the target which will be an absolute measure of their relative marksmanship, we may divide the sum of the squared radii (r^2) by the number of shots (N) in order to get the average squared error or miss. The square root of this will be the best possible measure of variability from the center of the target taken as a mean.

Another point in favor of the use of standard deviation is the fact that it bears a definite relation to the normal probability curve, or curve of error. It bears the same relation to the curve that the radius of a circle bears to the circle. It is therefore a constant and limits the *spread of the curve* so that, when the standard deviation is small, the measures are concentrated near the center, and the curve rises rapidly, whereas, if the standard deviation is great, the curve is flat, and the measures are scattered widely from the center. As stated above, it is a constant and marks the point where the curvature of the curve changes from the convex to the concave as you go from the center. Its value includes 68.26 per cent of the cases.

The Coefficient of Variability.—In the foregoing pages we have noted the principal methods of representing absolute

variability of frequency distributions. With the exception of the cases of variability in reference to marksmanship no reference was made to the relative variabilities of two or more distributions. The emphasis was rather to comprehend more fully the distribution of the measures in reference to their measure of central tendency. To measure the amount of "scatteration," or spread from the measure of central tendency, we employed these measures: (1) mean deviation, (2) standard deviation, and (3) probable error. It would be a simple matter to compare the variability of one distribution with another by the foregoing methods, since the units of variability were the same; that is, in the case of grades they were recorded in per cent and in the case of marksmanship they were recorded in inches. It very frequently happens, however, that we desire to compare the variability of one distribution with another when the units are different, as, for instance, comparing the salaries of teachers with the length of time they have been in the service. In a distribution of this kind one of the units would be dollars and the other years. It is therefore necessary to devise a measure of *relative variability* to cover these cases.

Pearson has devised such a measure, which he calls the *coefficient of variation*. It is the measure of the ratio of absolute variability (standard deviation, mean deviation, quartile deviation, or probable error) to the average from which these deviations are taken (arithmetic mean, or median.) It may be expressed by the formula

$$V = \frac{100\sigma}{M},$$

in which V is the measure of variability, σ is the standard deviation, and M the median, or mean. By this measure one is merely finding the per cent that the absolute variability bears to the average from which the deviations are

computed. It is clear that any other measure of variability might be used instead of the standard deviation.

A measure of this kind is independent of the units used and will show relative variability even though the units in the two distributions compared are entirely different. Thorndike proposes to take the square root of the measure of central tendency instead of using it as did Pearson. His formula would read:

$$V = \frac{100 \text{ } M.D.}{\sqrt{\text{Median}}}$$

The Pearson coefficient of variability seems to be a better measure, however, taking one distribution with another, than the one proposed by Thorndike.

CHAPTER XII

THE MEASUREMENT OF RELATIONSHIP, OR CORRELATION

Need for Measures of Relationship.—One more group of measures is necessary to equip the student with adequate means for dealing with educational data. We desire not only to know the distribution of measures in a series of educational data, but many times we desire to compare one series with another and therefore need some measure that takes cognizance of the distribution of the various cases in the series and at the same time expresses the movements of the group as a whole.

It is sometimes stated that there is a high correlation between abilities in mathematics and abilities in Latin. In order to compare two groups of this kind it is necessary to have measures that express the efficiency of each group as a whole. It is sometimes necessary to know how one group varies as compared with another. The comparison of one series of data with another is usually spoken of as correlation.

The measurement of relationship or correlation is a technical thing. Its relation to causation is also technical and difficult to understand. In order to help clarify the subject and present it from a number of points of view, we shall refer to the statement of the problem as discussed by a number of the leading statisticians who have written on the subject. We shall also give a number of simple illustrations which will throw some light on the solution of the problem.

Comparison involves the pairing of things or events that

are not identical in all particulars as to time, place, and condition. A study of cause and effect, whether coincidence or sequence, becomes largely a study of association. Causes never operate twice under exactly the same circumstances. Oneness of effect is only apparent. When making comparisons in education or psychology, there is a tendency to attempt to safeguard oneself against error and criticism by introducing the proviso, "*other things being equal.*" But "*other things*" are rarely, if ever, equal in actual life. Pearson says:¹

Individual phenomena can only be classified and our problem turns on how far a group or class of like, but not absolutely same, things which we term "causes" will be accompanied or followed by another group or class of like, but not absolutely same, things which we term "effects."

Bowley discusses correlation as follows:²

When two quantities are so related that the fluctuations in one are in sympathy with fluctuations in the other, so that an increase or decrease of one is found in connection with an increase or decrease (or inversely) of the other, and the greater the magnitude of the changes in the one, the greater the magnitude of the changes in the other, the quantities are said to be correlated.

Davenport says:³

The whole subject of correlation refers to that interrelation between separate characters by which they tend, in some degree at least, to move together. This relation is expressed in the form of a ratio.

In reference to zero correlation he says:

If the characters in question are absolutely indifferent the one to the other, the correlation is said to be zero, indicating mere association under the law of independent probability, without causative relation of any kind.

¹ Karl Pearson, *The Grammar of Science*, p. 157.

² A. L. Bowley, *Elements of Statistics*, p. 316.

³ *Principles of Breeding*, p. 453.

Pearson says:⁴

When we vary the cause, the phenomena changes, but not always to the same extent; it changes, but has variations in its change. The less the variation in that change, the more nearly the cause defines the phenomena, the more closely we assert the association or correlation to be. It is this conception of correlation between two occurrences embracing all relationships from absolute independence to complete dependence, which is the wider category by which we have to replace the old idea of causation. Everything in the universe occurs but once; there is no complete sameness of repetition.

Secrist says:⁵

A measure of correlation is a statement of probabilities, the reliability of which is determined by the degree to which the samples represent the whole "population" and the conditions under which the samples are taken, the range of condition.

We are striving to devise some methods of determining the degree of causal connection exhibited by certain traits and activities in school work. The measuring of mental, social, and physical activities constantly involves the study of causation, or causal connection, between two or more traits in question.

One of the psychological problems about which there has been much discussion in the last decade is the question as to whether or not "school" abilities are specialized or general. For example, what is the probability that a pupil who shows a high degree of achievement in Latin will show a high degree of achievement in mathematics? If we had several hundred cases and we found that some students who were good in Latin were also good in mathematics, that some who were good in Latin were mediocre in mathematics, and that a few who were good in the subject of

⁴ *The Grammar of Science*, p. 157.

⁵ *An Introduction to Statistical Methods*, p. 438.

Latin were very poor in mathematics, and vice versa, it would be very difficult to draw a conclusion as to the correlation between the two traits unless we had some way of measuring the amounts of correspondence between them.

Or, again, we might have a group of data showing that the student who ranked first in Latin also ranked first in mathematics, and the student who ranked second in Latin ranked second in mathematics, and so on. Data of this kind would show that there is a causal relation existing between the two traits, but it would not show how much. This method of measuring the degree of correspondence between two traits *takes account only of the position or rank of the various measures in the series and neglects the absolute amounts of the measures*. For the measurement to be complete the actual proportional differences between each two consecutive marks must be measured.

Two methods have been devised that take cognizance of the actual size of the various measures of the traits and show the degree of correspondence between them. The first is the graphic method, and the second is by the use of mathematical formulas. While the first method does take cognizance of the size of each measure in the distribution, it nevertheless must be refined by the application of mathematics before the absolute value may be found.

We shall first show how to represent the degree of correlation between two traits by the use of mathematical formulas, and then present methods for representing it graphically. It may seem to be more logical to present a graphic representation of the subject first, but it is believed that the subject may be made clearer by reversing what might seem to be the logical order.

Perhaps enough has been said about correlation and causation to introduce the reader to the subject. We shall now illustrate the computation of the coefficient of correlation

by a number of simple problems and supplement the previous treatment by additional discussion as new problems arise.

1. Illustrating the Computation of the Coefficient of Correlation: Data Simple and Ungrouped.—The unit with which we generally measure the degree of likeness or correlation of one series with another is called the *coefficient of correlation*. The formula used is that derived by Karl Pearson and is variously expressed as follows:

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}} \text{ or } r = \frac{\Sigma xy}{N \sigma_x \sigma_y}$$

The Meaning of Symbols Used

- r = the coefficient of correlation.
- Σ = the summation of the x -deviations multiplied by the corresponding y -deviations. (Σ always means summation when used in the formulas.)
- x = the deviation of each particular measure from the average in the first, or subject series.
- y = the deviation of each particular measure from the average in the second, or relative series. In representing the deviations by x and y the proper algebraic signs must be retained.
- x^2 = the square of the deviation of the subject series.
- y^2 = the square of the deviation of the relative series.
- σ_x = (second formula) equals the standard deviation of the first series and is the same as the $\sqrt{x^2}$ in the first formula.
- σ_y = the standard deviation in the second series.
- N = the number of pairs of items in the series.

The coefficient of correlation r is a constant and a measure of great importance in expressing correlation. It is evidently a pure number, and its magnitude is unaffected by the units in which x and y are measured; for the numerator and denominator are affected to the same extent.

The development of the Pearson formula gives values

for r ranging from -1 through 0 to $+1$. If $r=+1$, the correlation is said to be perfect and positive and means that large values in the first series are accompanied by large values in the second series, and vice versa. If $r=-1$, the correlation is perfect but negative and means that large values in the first series are accompanied by small values in the second series, and vice versa. When $r=0$ the two series are independent.

TABLE XXXV.—ILLUSTRATING THE COMPUTATION OF THE COEFFICIENT OF CORRELATION BETWEEN ADDITION AND HANDWRITING

(Hypothetical Case)

Pupils	Scores in Addition (Subject Series)	Scores in Handwriting (Relative Series)
A	3	5
B	4	8
C	5	7
D	8	12
E	10	13
	5)30	5)45
	6 = mean	9 = mean

The deviation of each score from its respective mean is given below; also the product of each deviation in the first series by its corresponding deviation in the second series.

Pupils	Deviations, x , from the Mean Subject Series	Deviations, y , from the Mean Relative Series	xy
A	-3	-4	12
B	-2	-1	2
C	-1	-2	2
D	+2	+3	6
E	+4	+4	16
			38

Squaring the individual deviations to find standard deviations we have:

x^2	y^2
9	16
4	1
1	4
4	9
16	16
<hr/>	<hr/>
34	46

We are now ready to substitute in the formula:

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}} = \frac{38}{\sqrt{34 \times 46}} = 0.962$$

In order to get a value for r equal to zero, it is evident that the numerator of the fraction in the above formula must be zero. A hypothetical case given in Table XXXVI will illustrate this point.

TABLE XXXVI.—ILLUSTRATING THE COMPUTATION OF THE COEFFICIENT OF CORRELATION EQUAL TO ZERO

First Series	Second Series	Deviations, First Series	Deviations, Second Series	xy
95	92	+25	+2	+50
85	90	+15	0	0
75	88	+5	-2	-10
65	88	-5	-2	+10
55	90	-15	0	0
45	92	-25	+2	-50
6)420	6)540			$xy = 0$
70 = mean	90 = mean			

Since $xy = \text{zero}$, therefore the correlation between the two series is zero.

TABLE XXXVII.—ILLUSTRATING A PERFECT POSITIVE CORRELATION

First Series	Second Series	Deviations, First Series	Deviations, Second Series	xy
20	30	-10	-25	250
24	40	-6	-15	90
28	50	-2	-5	10
32	60	+2	+5	10
36	70	+6	+15	90
40	80	+10	+25	250
6)180	6)330			$xy = 700$
30 = mean	55 = mean			

x^2	y^2
100	625
36	225
4	25
4	25
36	225
100	625
280	1750

Substituting formula,

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}}$$

$$\frac{700}{\sqrt{280 \times 1,750}} = \frac{700}{700} = 1$$

Therefore the correlation is perfect. A perfect negative correlation may be illustrated in the same way.

2. Illustrating the Computation of the Coefficient of Correlation: Data Complex and Grouped in Class Intervals.—In Tables XXXV, XXXVI, and XXXVII we illustrated the computation of the coefficient of correlation by simple problems where the number of pairs of values were five and six. It is evident that, if we had hundreds of pairs of values and attempted to use the same procedure, the amount of labor in the computation would be very great. A glance at the Pearson formula shows that the coefficient of corre-

lation is found in terms of measures of central tendency and measures of variability; that is, we compute the mean of each series, find the deviation of each measure from its mean, multiply it by the corresponding deviation from the mean in the other series, and find the sum of these products for the numerator of the fraction. The denominator is the product of the standard deviations of the two series. There is therefore really nothing new in the development of this formula, since we showed in the two previous chapters how measures of central tendency and variability might be found. The chief concern here is to devise a distribution table that will reduce the labor to a minimum. Two plans will be presented.

One will be a double entry distribution table which is simply a device for the arrangement of the pairs of values to facilitate computation, and the other will be a short method for computing correlations proposed by Dr. Leonard P. Ayres.⁶

We shall first note the computation of the coefficient of correlation by the traditional method, using a double entry distribution table, and at the same time take advantage of the short methods in computing the means and deviations developed in Chapters IX and X.

Table XXXVIII illustrates the tabulation of the data and the computation of the coefficient of correlation with 310 cases in the subjects of Latin and mathematics. In this table the mathematics is known as the *first* or *subject* series, and the Latin, the *relative* or *second* series. The columns and rows are spoken of as *arrays*, the columns as *y-arrays of type x*, and the rows as *x-arrays of type y*. The

⁶ "A Shorter Method for Computing the Coefficient of Correlation," *Journal of Educational Research*, Vol. 1, March, 1920. Also "The Application to Tables of Distribution of a Shorter Method for Computing Coefficients of Correlation," *Journal of Educational Research*, Vol. 1, April, 1920.

size of the class intervals is determined here in the same way as it was in the previous discussions in Chapters IX and X. There are five units in each class interval. The limits of the class intervals are written at the upper and left-hand margin of the table. The number of class intervals is determined by the range of the grades; from 40 to 100 in Latin, and from 60 to 100 in mathematics. Noting the table (first row at the top) we see that 5 students made grades of from 95 to 100 in Latin and from 65 to 70 in mathematics; 20 made grades of from 95 to 100 in Latin and from 80 to 85 in mathematics, and so on.

The heavy black horizontal lines marking the limits of the class interval 70-74.99 designate the row and class interval that contains the assumed mean of the Latin series. The heavy black vertical lines at the limits of the class interval 80-84.99 mark the column and class interval that contains the assumed mean for the mathematics series. The small figures in the upper right-hand corner of the squares are the deviations from the assumed mean (in units of class intervals) of the subject series, multiplied by the corresponding deviations from the assumed mean of the relative series, and this product multiplied by the number of measures in the class interval. Thus the -75 in the second column from the left and the first row from the top means that the 5 measures in that square have a deviation of -3 from the assumed mean of the mathematics series and a +5 from the assumed mean in the Latin series. Hence: $-3 \times 5 \times 5 = -75$. The other figures are determined in a similar way. The fx -row at the bottom gives the sum of the measures in the various columns, and the fy -column at the right gives the sum of the measures in the various rows. The $\Sigma x'y'$ -column is the sum of the products of the deviations of each measure from the assumed mean in each series; that is, $-75 + 70 = -5$; $-72 + -96 = -168$, and so on.

TABLE XXXVIII.—ILLUSTRATING THE COMPUTATION OF THE COEFFICIENT OF CORRELATION OF 310 CASES OF LATIN AND MATHEMATICS

Ability in Mathematics													
	60	65	70	75	80	85	90	95	f_y	d	fd_y	$f d_y^2$	$\Sigma x'y'$
100													
95		-75			0		70		32	5	160	800	- 5
94.99					20		7						
90		-72	-26							18	72	288	-168
89.99		6	12										
85					0		45		40	3	120	360	45
84.99					35		5						
80						40	45		32	2	64	128	68
79.99					20	12							
75				-11	11				11	1	11	11	- 11
74.99					0	0					427		
70					79	1			80	0			
69.99													
65									0	-1			
64.99		84		24									
60		14		12					26	-2	- 52	104	108
59.99													
55		108				-81			39	-3	-117	351	27
54.99		12				27							
50									2	-4	- 8	32	32
49.99													
45									0	-5			
44.99													
40					0				30	-6	-180		
f_x	2	37	12	102	86	47	19	5	310		-357		
												1,080	
												310)3,154	310)116
												10.174 = s_y^2	0.374 = $\frac{\Sigma x'y'}{N}$
												0.051 = c_y^2	
												10.123 = σ_y^2	
												3.181 = σ_y^2	

d 4 3 2 1 0 1 2 3

$f d$ 8 111 24 102 245 47 38 15 100 246 310)145 0.468 = c_x^2 0.219 = c_x^2

$f d^2$ 32 333 48 102 47 76 45 15 100 246 310)683 2.203 = s_x^2 0.219 = c_x^2 1.984 = σ_x^2 1.408 = σ_x^2

427
310) 70
0.226 = c_y^2
0.051 = c_y^2
 $c_x c_y = -0.105$
 $\sigma_x \sigma_y = 4.479$
 $r = \frac{0.374 - (-0.105)}{4.479} = \frac{0.479}{4.479} = 0.107$ Ans.

SUMMARY OF STEPS IN THE COMPUTATION OF THE COEFFICIENT OF CORRELATION

1. Find the number of measures in each series and designate the number by N . (N will, of course, be the same in each series.)

2. Estimate the class interval that contains the mean and take its mid-point as the assumed mean (the same as was done in finding the mean by the short method in Chapter IX). For example, 70-74.99 for the y 's and 80-84.99 for the x 's.

3. Tabulate the deviations from the means chosen in terms of class intervals. For example, the first row above the class interval 70-74.99 has a deviation of +1, the second row +2, and so on. The first row below has a deviation of -1, the second a deviation of -2. The first column to the right of the mean column 80-84.99 has a deviation of +1, the second, +2; the first to the left a deviation of -1, and the second a deviation of -2, and so on.

4. To the right of the table make a d -column to record the y -deviations, also an fd_y -column, an fd_y^2 -column and an $\Sigma x'y'$ -column, which latter is the sum of the products of x and y deviations calculated from an assumed mean. Similarly record the x deviations and other similar values at the bottom of the table.

5. Multiply each frequency (fy -column for the y 's) by its respective deviation. For example, $32 \times 5 = 160$; $18 \times 4 = 72$; and so on, retaining the algebraic signs. Find the fd 's for the x 's in a similar way.

6. Find the algebraic sum of the fd_y 's. For example, $\Sigma fd_y = -357 + 427 = 70$; similarly $\Sigma fd_x = -245 + 100 = -145$.

7. Divide the Σfd 's by the number of cases, N , to give the correction c . For example, $c_y = \frac{70}{310} = 0.226$; $c_x = -\frac{145}{310} = -0.468$.

8. Square the corrections $c_y^2 = 0.051$; $c_x^2 = 0.219$.

9. Multiply each fd_y by d , its corresponding deviation, giving the figures in the fd_y^2 -column. For example, in the y -series, $160 \times 5 = 800$; $72 \times 4 = 288$, and so on. In the x -series $-8 \times -4 = 32$; and $-111 \times -3 = 333$.

10. Find the sum of the fd^2 's in each series. For example, $\Sigma fd_y^2 = 3,154$, and $\Sigma fd_x^2 = 683$.

11. Divide each of those sums by N , the number of cases, to give S^2 , the square of the standard deviation of each distribution around the assumed mean. $S_y^2 = \frac{3,154}{310} = 10.174$; $S_x^2 = \frac{683}{310} = 2.203$.

12. Subtract c_y^2 , the square of the correction, from S_y^2 , and c_x^2 from the square of the correction S_x^2 . For example, $10.174 - 0.051 = 10.123 = \sigma_y^2$; $2.203 - 0.219 = 1.984 = \sigma_x^2$.

13. Find the square roots of σ_y^2 and σ_x^2 . For example, $\sigma_y = 3.181$; $\sigma_x = 1.408$.

14. Compute the $\Sigma x'y'$'s by finding the sum of the deviations of the measures in a particular row from the mean of the x 's of the whole table, \bar{x} . This gives $\Sigma x'$. Multiply $\Sigma x'$ by y' , the deviation of this particular row from \bar{y} , the mean of the y 's of the whole table. This gives $\Sigma x'y'$, which is the product-sum of the deviations about the two assumed means.

This would give us the numerator of the fraction in the Pearson formula were it not for the fact that both the x' and y' deviations have been computed from assumed means instead of the true means and hence must be corrected. Since the means were in error by the corrections c_x and c_y , so each deviation from them on y and x must be in error by the same amount.

Going through the various rows and finding the algebraic sum of the products of $x'y'$ we get -5 , -168 , 45 , etc., of the $\Sigma x'y'$ -column. For example, adding -75 and $+70$ gives -5 ; -72 and -96 gives -168 , and so on. The algebraic sum of the $\Sigma x'y'$ -column = 116 , which, when divided by the total number of measures, $N = \frac{116}{310} = 0.374 =$

$\frac{\Sigma x'y'}{N}$; $c_x c_y = 0.226 \times -0.468 = -0.105$. It may be shown algebraically that Σxy , the deviations of x and y from their true means $= \frac{\Sigma x'y'}{N} - c_x c_y$.

We may therefore write the Pearson formula thus:⁷

⁷ Let E_x and E_y represent the estimated means of the two series and c_x and c_y be corrections to be applied to the estimated means to get the true means. Then the true means, M_x and M_y , are respectively, $M_x = E_x + c_x$ and $M_y = E_y + c_y$.

Let x and y be deviations from the true means, M_x and M_y .

Let x' and y' be deviations from the estimated means, E_x and E_y .

$$r = \frac{\Sigma x'y' - N c_x c_y}{N \sigma_x \sigma_y} = \frac{\frac{\Sigma x'y'}{N} - c_x c_y}{\sigma_x \sigma_y}.$$

We now have made all corrections so that we may substitute directly in the formula,

$$r = \frac{\frac{\Sigma x'y'}{N} - c_x c_y}{\sigma_x \sigma_y}$$

which equals the Pearson formula,

$$r = \frac{\Sigma xy}{N \sigma_x \sigma_y}$$

Substituting,

$$r = \frac{0.374 - (-0.105)}{3.181 \times 1.408} = \frac{0.479}{4.479} = 0.107.$$

3. Illustrating the Computation of the Coefficient of Correlation by the Short Method (Adapted from Ayres): (a) Series Simple and Ungrouped.—A shorter and more direct

Thus

$$x' = x + c_x \quad \text{and} \quad y' = y + c_y.$$

Therefore,

$$\begin{aligned} \Sigma x'y' &= \Sigma (x + c_x)(y + c_y) \\ &= \Sigma xy + c_y \Sigma x + c_x \Sigma y + \Sigma c_x c_y. \end{aligned}$$

Now since Σx and Σy (the sum of the x and y deviations from the true mean) each = 0, then

$$\Sigma x'y' = \Sigma xy + \Sigma c_x c_y, \quad \text{or} \quad \Sigma xy = \Sigma x'y' - \Sigma c_x c_y.$$

or, substituting this expression in the equation,

$$r = \frac{\Sigma xy}{N \sigma_x \sigma_y},$$

we get,

$$r = \frac{\Sigma x'y' - N c_x c_y}{N \sigma_x \sigma_y} = \frac{\frac{\Sigma x'y'}{N} - c_x c_y}{\sigma_x \sigma_y}.$$

(Adapted from H. L. Rietz, Bulletin No. 148, University of Illinois Agricultural Experiment Station, 1910.)

method has been evolved by Ayres.⁸ We shall now illustrate the computation of the coefficient of correlation by this method by two simple series with data ungrouped. The method used for finding the coefficient of correlation in Table XXXVIII was a long one even though many "short cuts" were used.

In Table XXXV we gave two simple series to illustrate the computation of the coefficient of correlation. The process will be repeated in part to show the advantages of the "short cuts" used by Ayres. The series were:

Subject Series	Relative Series
3.....	5
4.....	8
5.....	7
8.....	12
10.....	13
<hr/>	<hr/>
30.....	45

The mean of the subject series is 6, that of the relative, 9. Calling the deviation from the mean in each series x and y respectively, and multiplying the deviation in one series by its corresponding deviation in the other and also finding the standard deviations of the two series, we have:

x	y	xy	x^2	y^2
-3	-4	12	9	16
-2	-1	2	4	1
-1	-2	2	1	4
+2	+3	6	4	9
+4	+4	16	16	16
		<hr/>	<hr/>	<hr/>
		38	34	46

which when substituted in the Pearson formula,

$$r = \frac{38}{\sqrt{34 \times 46}} = 0.962$$

The shorter method proposed by Ayres gives the sums of the products and the sums of the squares of the deviations

⁸ Ayres, *loc. cit*

directly from the squares of the original numbers. By this method it is not necessary to arrange the measures in order of their magnitude or to find the separate deviations from the means. Thus is avoided the necessity of taking into account the plus and minus signs of deviations.

There are two fundamental principles in the computation of the coefficient of correlation by this method which the student should master.

1. This method considers *every number in the series as being equal to the mean of the series plus a plus or minus deviation from that mean.* Thus 3, which is the first number in the first series in the above illustration, equals the mean of the series, or 6, plus a minus deviation, -3 ; 4 equals the mean of the series, or 6, plus a minus deviation, -2 ; and so on.

2. *If the sum of the squares of the numbers in a series is found and from it is subtracted the product found by multiplying the square of the mean of the series by the number of cases, the remainder will be the sum of the squares of the deviations from the mean.*

For example, in the subject series given above the numbers, their means, and squares are:

	3 squared =	9	
	4 " =	16	
	5 " =	25	
	8 " =	64	
	10 " =	100	
	5)30	214	
Mean =	6	180	
Mean squared =	36		34 the sum of the squares of the de-
Number of cases =	5		viations of the subject items.
		180	

By utilizing this method the coefficient of correlation may be computed directly from the products of the squares of the items of the two series without finding the separate deviations.

The operations to be performed may be expressed as follows:

$$r = \frac{\Sigma(S \times R) - \frac{\Sigma S \times \Sigma R}{N}}{\sqrt{\left(\Sigma S^2 - \frac{(\Sigma S)^2}{N}\right) \left(\Sigma R^2 - \frac{(\Sigma R)^2}{N}\right)}}$$

where S = the individual subject items as 3, 4, 5, 8, 10, in the subject series above

R = the individual relative items

Σ = the sum

S^2 = the square of each individual subject item, as 9, 16, 25, 64, 100

R^2 = the square of each individual relative item

N = the number of cases

Making the corrections and substituting in the formula with data used above, we have

$$\begin{aligned} r &= \frac{308 - \frac{30 \times 45}{5}}{\sqrt{\left(214 - \frac{900}{5}\right) \left(451 - \frac{2,025}{5}\right)}} \\ &= \frac{308 - 270}{\sqrt{(214 - 180)(451 - 405)}} \\ &= \frac{38}{\sqrt{34 \times 46}} = \frac{38}{39.5} = 0.962 \end{aligned}$$

In order to make the above substitutions clear, let us go through the various steps and note how each part is derived. The 308 is the sum of the products of the subject and relative items, that is: $3 \times 5 + 4 \times 8 + 5 \times 7 + 8 \times 12 + 10 \times 13 = 308$.

The fraction $\frac{30 \times 45}{5}$ is the sum of the subject items, 30, multiplied by the sum of the relative items, 45, and divided by the number of cases, 5. The first number under the

radical sign, 214, is the sum of the squares of the items in the subject series. The fraction, $\frac{200}{5}$, does not seem at first to carry out the directions in the second principle stated above which says to subtract from the sum of the squares of the numbers the product found by multiplying the square of the mean of the series, 36, by the number of cases, or 5. The mean is 6, or $\frac{30}{5}$, which when squared = 36, or $\frac{200}{5}$, and when multiplied by the number of cases gives 180, or $\frac{200}{5}$. The other numbers under the radical are found in the same way.

(b) *Data Complex and Grouped in Class Intervals.*—We shall now illustrate the computation of the coefficient of correlation with the data grouped in class intervals. To show the similarity of this method to the one used in Table XXXVIII we shall use the same data that were used in that table.

The data should be arranged in a correlation table the same as in Table XXXVIII. At the left of the table (see Table XXXIX), and at the bottom, instead of inserting the class intervals 40 to 44.99, 45 to 49.99, etc., as was done in Table XXXVIII, insert the numbers 1, 2, 3, etc., in the column marked *S*. In order to keep the data straight, the class intervals may be inserted if desired as in Table XXXVIII and the numbers in column *S* inserted later. It should be noted that multiplying the frequencies by the numbers 1, 2, 3, etc., gives them the same relative values as if they were multiplied by the mid-points of the class intervals 42.5, 47.5, 52.5, etc. In like manner insert the figures 1, 2, 3, etc., in the margin at the top of the table. When data are grouped in class intervals, we may assume that they are grouped at the mid-point of the class interval. Thus, in the first row at the top, the class interval is 95 to 100 (see Table XXXVIII), and all measures are assumed to be grouped at the mid-point, or at 97.5. The class interval in the second row is 90 to 94.99, and its mid-point is 92.5. Therefore all measures in the top row may be said to have a

value of 97.5 and in the next row, 92.5, and so on. Similarly the measures in the columns may be considered as grouped at their mid-points as in Table XXXVIII. The first column to the left would then have a value 62.5, the second one 67.5, and so on. To compute the coefficient of correlation by this method it is necessary to multiply the total number of measures in the rows and columns by their magnitude. Thus, to get the *ST*-column at the right of the table, we accordingly would have to multiply the total number of measures in the first row, summed in the *T*-column, by the magnitude of the measures, which in this case would be 97.5. In order to get the measures in the *SST*-column we would have to square the 97.5 and multiply it by the sum of the measures in the first row, or 32. But in order to avoid the squaring of large numbers and also the multiplication by large numbers, we insert the numbers 1, 2, 3, etc., instead of the mid-values of the class intervals.

The *T*, *RT*, and *RRT* rows at the bottom are found in the same way as the *T*, *ST*, and *SST* columns at the right. (Use Table I, Appendix, for squaring the numbers.)

The ΣSf -row at the bottom of the table is found by multiplying the frequencies *f* by the subject series, *S*, finding their sum and recording in the proper squares below. Thus in the first column $3 \times 2 = 6$; in the second column, $12 \times 5 = 60$; $11 \times 6 = 66$, $5 \times 14 = 70$; $4 \times 12 = 48$; $60 + 66 + 70 + 48 = 244$. The other numbers in the row are found in a similar way.

The *R* (ΣSf)-row is found by multiplying the numbers in the ΣSf -row by *R*, the relative series. Thus $1 \times 6 = 6$; $2 \times 244 = 488$; $3 \times 132 = 396$, etc.

In the computation below the diagram the numbers are taken from the totals at the right and the bottom of the columns and rows respectively. Substituting the values in the Pearson formula we have, $r = 0.107$, the same as by the other method on page 334.

MEASUREMENT OF CORRELATION 343

TABLE XXXIX.—ILLUSTRATING THE COMPUTATION OF THE COEFFICIENT OF CORRELATION BY THE SHORT METHOD

(Adapted from Ayres)

		Subject Series										
Relative Series	S \ R	1	2	3	4	5	6	7	8	T	ST	SST
	12		5			20		7		32	384	4,608
	11		6	12						18	198	2,178
	10					35			5	40	400	4,000
	9						20	12		32	288	2,592
	8				11					11	88	704
	7				79	1				80	560	3,920
	6									0	0	0
	5		14		12					26	130	650
	4		12				27			39	156	624
	3	2								2	6	18
	2									0	0	0
	1					30				30	30	30
	T	2	37	12	102	86	47	19	5	310	2,240	19,324
	RT	2	74	36	408	430	282	133	40	1,405	2,240 310	= 7.226
	RRT	2	148	108	1,632	2,150	1,692	931	320	6,983	1,405 310	= 4.532
	ΣSf	6	244	132	701	627	288	192	50			
	R(ΣSf)	6	488	396	2,804	3,135	1,728	1,344	400	10,301		

$$\begin{aligned}
 2,240 \times 4.532 &= 10,151.68 & 10,301 - 10,151.68 &= 150 \\
 2,240 \times 7.226 &= 16,186.24 & 19,324 - 16,186.24 &= 3,138 \\
 1,405 \times 4.532 &= 6,367.46 & 6,983 - 6,367.46 &= 616 \\
 r &= \frac{150}{\sqrt{3,138 \times 616}} = 0.107
 \end{aligned}$$

We may further clarify the process by comparing the steps here with the simpler problem solved by the Ayres short method on page 340. The fraction $\frac{2240}{310} = 7.226$, the mean of the relative series. $\frac{1405}{310} = 4.532$, the mean of the subject series. It should be noted that in each case it is the sum of the values of the individual measures divided by the number of cases. $10,301 = \Sigma(S \times R)$ in the general formula and corresponds to the 308 in the simple problem. $2,240 \times 4.532 = 10,151.68 = \frac{\Sigma S \times \Sigma R}{N}$ of the general formula;

the 2,240 being ΣS and the 4.532 being the $\frac{\Sigma R}{N}$. $10,301 - 10,151.68 = 150$, the simplified numerator of the fraction corresponding to 38 in the simple problem. $6,983 = \Sigma S^2$ in the denominator of the general formula and $19,324 = \Sigma R^2$; $2,440 \times 7.226 = 16,186.24 = \frac{\Sigma R^2}{N}$ and $1,405 \times 4.532 = 6,367.46 = \frac{\Sigma S^2}{N}$.

4. Representing the Degree of Correlation Between Two Traits by the Graphic Method: (a) *Data Simple and Ungrouped*.—In making a correlation table and inserting the values we must, of course, deal with each pair of correlated values separately in order to locate them in the right place in the correlation table. In plotting pairs of measures we construct two coördinate axes, one horizontal (OX), and the other vertical (OY), meeting at an origin, or beginning point, at the bottom and left. On these axes we lay off scales representing the traits in question. The scales need not be made up of the same units. If we desire to find the correlation between the heights and weights of individuals, for instance, we would lay off one scale on, say, the OX -axis, to represent height and the other on the OY -axis to represent weight. The units of the scale are laid off from the origin to the right on OX and upward on OY .

When one desires to insert a pair of measures in a correlation table thus constructed, he finds the proper place on the scale on the X -axis for one trait and the proper place on the Y -axis for the corresponding trait. He then erects perpendiculars at each of the points thus located and at the intersection of these perpendiculars a dot or small cross is made, which represents the pair of values in the correlation

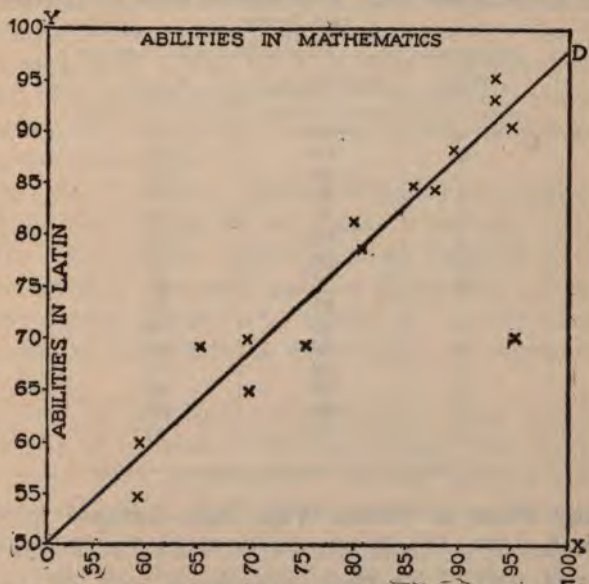


FIGURE X. ILLUSTRATING THE DISTRIBUTION OF CORRELATED ABILITIES IN LATIN AND MATHEMATICS

table. In Table XL below are given the grades made by 15 pupils in mathematics and Latin. Figure X shows the location of these 15 pairs of values in a correlation table. Pupil A made a grade of 95 per cent in mathematics and 93 per cent in Latin. In inserting these values we find the place on the X -axis that is marked 95 and erect a perpendicular to the X -axis at that point. Similarly, we find the

point on the Y-axis that corresponds to a grade of 93 per cent and erect a perpendicular to the Y-axis at that point. At the intersection of these two perpendiculars we make a small cross which represents the pair of values in the correlation table. The other grades are similarly located in the table.

TABLE XL.—GRADES MADE BY FIFTEEN PUPILS IN MATHEMATICS AND LATIN

Pupil	Mathematics	Latin
A	95	93
B	93	93
C	90	94
D	88	89
E	87	85
F	70	95
G	80	82
H	86	85
I	81	79
J	75	70
K	70	65
L	69	71
M	65	69
N	60	60
O	55	60

Many Pairs of Values With Data Grouped in Class Intervals.—If perpendiculars were erected at the junction points of each of the class intervals on both the X- and Y-axes in Figure X, we would have a convenient device for tabulating any number of pairs of values in the correlation table. When data are thus grouped in class intervals, we make the same assumption we did in the computation of the mean and median, namely, that all the measures are grouped at the mid-point of the class interval. Considering any one square in the distribution table we assume that the measures are grouped at the mid-point of the class interval for Latin abilities and also for mathematics

abilities, hence all the measures in a square are considered to be grouped at the mid-point of the square. When the data are thus recorded in the correlation table we count the number of measures in each square and insert the proper figure to represent these measures.

We are now ready to make an estimate of the degree of correlation existing between the two traits. This may be done by drawing a line that most closely approximates the general scattering of the pairs of measures over the table. In drawing such a line we must take cognizance of the number of measures in each square. The line OD is the best fitting line for the 15 pairs of values in Table X and is known as the correlation line.

It is evident that this method is inaccurate, because we could not take two similar correlation tables and tell which had the higher degree of correlation, for the position of the line OD is only estimated and not determined accurately. The measures might be rather similar in "scatter" but one would be unable to tell the exact degree of correlation in either table.

In order to represent both graphically and accurately the degree of correlation existing between two traits, we must resort to a mathematical formula, such as the one devised by Pearson, that will take cognizance of the exact value of each measure in the correlation table. A line drawn that most closely approximates the general scattering of the pairs of measures in the table will bear a constant relation to the X -axis, which is expressed by the fraction $\frac{y}{x}$.

But the ratio $\frac{y}{x}$ is the tangent of the angle made by the correlation line and the X -axis (measured from the Y -axis if the angle is more than 45 degrees.) We may take, therefore, any value of $\frac{y}{x}$ (which will be a value between 0 and 1)

and with a table of natural tangents determine the number of degrees the line of correlation makes with the X -axis. The line may then be drawn accurately in the correlation table.

We may further illustrate the graphic method of expressing correlation by referring to Figure XI. In that figure let the line XOX' be the mean of abilities in mathematics and the line $Y'OY$ the mean of the abilities in Latin. Then

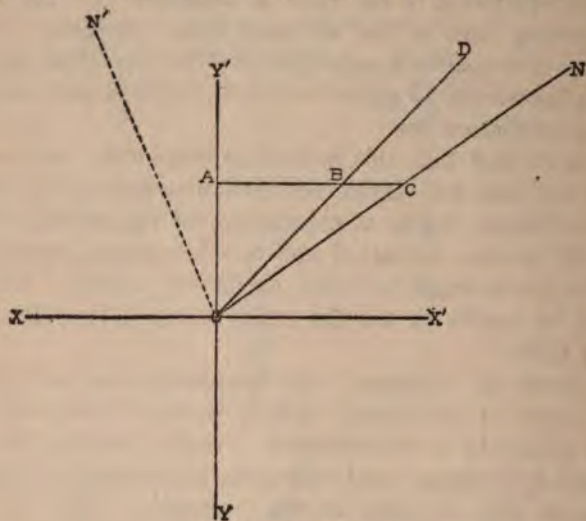


FIGURE XI

it is evident that if a student who was 5 units above the mean in Latin was also 5 units above the mean in mathematics, and that if one 3 units above the mean in Latin was 3 units above in mathematics, and the same ratio prevailed for all students both above and below the means of the two abilities, a curve plotted from these data would be a straight line and would bisect the angle $Y'OX'$; that is, it would make an angle of 45 degrees with the X -axis, and the correlation would be perfect. If, however, the correlation were not

perfect, then the line would make some other angle with the X -axis, depending on the degree of correlation between the two traits.

If the correlation were zero, the line would take the position OY' or OX' . This would mean that any given change in a y -value would be accompanied by no change in an x -value and vice versa. If the correlation were negative, then the line would swing to the left of the line OY' taking any position as ON' . This would mean that as the y -values grow larger the x -values would grow smaller. If we have a perfect negative correlation, the line ON' would make an angle of 45 degrees with the X -axis.

It thus may be seen then that the degree of correlation existing between two traits is expressed by the angle that the line of correlation makes with the X -axis. If the x -values increase more rapidly than the y -values, the line of correlation will lie in the angle DOX' . If the y -values increase more rapidly than the x -values, it will lie in the angle DOY' . It should be noted that, since the lines XOX' and YOY' represent the means of the two traits, any x -value greater than the mean will lie to the right of line, YOY' and will deviate from the mean positively, while a value below the mean will lie to the left of the line YOY' and will be a negative deviation. In like manner any y -value above the mean will have a positive deviation from the mean and will lie above the XOX' line, while a value less than the mean will lie below XOX' . The signs for the four quadrants may, therefore, be represented thus:

<i>Second Quadrant</i> $x = -$ $y = +$	<i>First Quadrant</i> $x = +$ $y = +$
<i>Third Quadrant</i> $x = -$ $y = -$	<i>Fourth Quadrant</i> $x = +$ $y = -$

When the correlation was not perfect, the degree of correlation was expressed by Galton thus: Draw any horizontal line AC cutting the line of perfect correlation OD at B and the line ON at C . Then the ratio $\frac{AB}{AC}$ measures the amount of correspondence in change in the two variables. A given change in the size of y is accompanied by a proportional change in the size of X . When the line ON swings to the position OD , then $\frac{AB}{AC} = 1$ and the correlation is perfect and positive. When ON takes the position OY , the ratio $\frac{AB}{AC} =$ infinity, since AC will equal zero; that is, a given change in the size of y is accompanied by no change in the size of x . When the line ON swings to the left of the line YOY' , the degree of correlation is measured in a similar way from that quadrant but the correlation is negative. The ratio $\frac{AB}{AC}$ is called the *coefficient of correlation* and is denoted by r .

We shall now note another very important term in the comparison of traits. Galton found in measuring the heights of individuals that if a group of parents were found to be, say y -inches above or below the mean of the race in stature that the mean stature of their children would not deviate y -inches from the mean of the race, but would deviate only $2/3y$ -inches above or below the mean of the race. In expressing this fact Galton said that the offspring tended to "regress" towards the mean of the race. Since then it has been common to speak of the line of means of the correlation table as the line of regression. Since there is a line of means of the columns and also one for the rows, there will, of course, be two lines of regression. The reader should note that the regression lines are not the same as the line that represents the coefficient of correlation. Each is expressed by a different equation, as will be shown later.

Finding the Equation of a Straight Line of Regression.

—The most definite way to describe a line is to write its equation. Since the relationship of most of the educational data may be expressed by a straight line, we shall note how to write the equation of a straight line. In order to do this we must be able to put two variables as x and y together in an algebraic equation in such a way that a given change in the value of one is accompanied by a proportional change

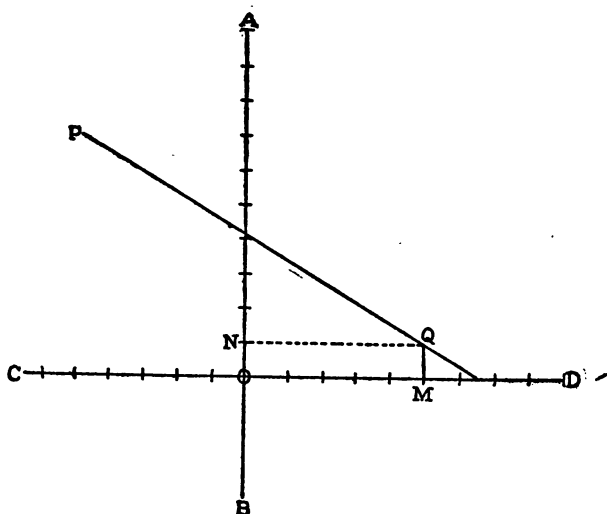


FIGURE XII

in the value of the other. Let us write the equation of the line PQ in Figure XII. The equation of a line may be written if we know the value of any two points on it. From Figure XII we note that point Q has an x -value of $+5$ and a y -value of $+1$. These values are measured from point O , the origin of the graph. The distance QM is called the *ordinate* of point Q and the distance QN is called the *abscissa* of point Q . The points $(5, 1)$ are called the *coordinates*

of Q , and in like manner $(-5, +7)$ are called the coördinates of P . The axes are spoken of as coördinate axes. The equation of the line PQ is $3x+5y=20$. For any two points on the line PQ , the ratio of the difference of their ordinates to the difference of their abscissas may be expressed as $\frac{y_1-y_2}{x_1-x_2}$. It is evident that this ratio is constant for all points on the line. This ratio is the tangent of the angle that the line PQ makes with the line CD . (The tangent of an angle is the side opposite the angle divided by the adjacent side.) Since the ratio $\frac{y_1-y_2}{x_1-x_2}$ measures the inclination of the line PQ , it is called the *slope* of the line.

Let point Q have an x -value of plus 5 and a y -value of plus 1 $(+5, 1)$ and point P have an x -value of minus 5 and a y -value of plus 7 $(-5, +7)$. Since point P has an x and y value and point Q also, the points may be designated $P(x_1, y_1)$ and $Q(x_2, y_2)$, x_1 and x_2 being the abscissas of points P and Q respectively, and y_1 and y_2 , the ordinates. The slope of the line PQ equals $\frac{y_1-y_2}{x_1-x_2}$. Let $P_1(xy)$ be any other

point on the line PQ . Then the slope of P_1P will be $\frac{y-y_1}{x-x_1}$, since P_1 , P , and Q are on one line, slope P_1P =slope PQ . Hence we have the formula $\frac{y-y_1}{x-x_1} = \frac{y_1-y_2}{x_1-x_2}$. Substituting in the equation the values for the coördinates of points P and Q , we have $\frac{y-7}{x+5} = \frac{7-1}{5-5} = \frac{y-7}{x+5} = -\frac{6}{10}$. Clearing fractions we have $3x+5y=20$, which is the equation of the line desired.

We now desire the equation of a line that will take cognizance of the "scatteration" of the measures from the means of the corresponding rows and columns. We noted that the degree of correlation is measured in terms of the

relative amounts of deviation of each point from the mean of the column and from the mean of the row in which it falls. In the measurement of dispersion from the means of the columns and the rows we must use the same unit of deviation, if we desire the amounts of dispersion to be comparable. The value of standard deviation having been demonstrated in Chapter XI, under the heading of "Measurement of Variability," as the best measure of dispersion, it is therefore employed to measure dispersion in writing the equation of the line of regression.

We know that the line that "best fits" the means of the arrays is that line from which the deviations of the means are the least possible. Professor Karl Pearson, who developed the equation of the line, employed the method of least squares (discussed in a previous chapter) in the location of this line; that is, he located the line in such a position that the sum of the squares of the deviations of the means of the arrays, each weighted by the number of measures in the respective arrays, would be the minimum.

Pearson's Equation for a Line of Regression.—Pearson deduced the equation of the "best fitting" line as:

$$y_1 - \bar{y} = r \frac{\sigma_y}{\sigma_x} (x_1 - \bar{x}).$$

Meaning of Symbols Used

\bar{y} = the mean of the columns

\bar{x} = the mean of the rows

y_1 = a particular measure in the columns

x_1 = a particular measure in the rows

r = the coefficient of correlation designated by $\frac{AB}{AC}$ in figure

σ_y = the standard deviation of the y series

σ_x = the standard deviation of the x series

The equation may be written in a condensed form as

$$y = r \frac{\sigma_y}{\sigma_x} x,$$

in which y = the deviation of a particular y -measure from its mean, or $y = y_1 - \bar{y}$. Similarly $x = x_1 - \bar{x}$.

We noted above that the slope of the line was represented by the ratio $\frac{y}{x}$ if the line passes through the origin of the graph. Let this value be represented by m ; that is, $m = \frac{y}{x}$ and $y = mx$. We also note that

$$y = r \frac{\sigma_y}{\sigma_x} x$$

Therefore,

$$mx = r \frac{\sigma_y}{\sigma_x} x, \text{ and } m = r \frac{\sigma_y}{\sigma_x}$$

The slope of the line is therefore represented by $r \frac{\sigma_y}{\sigma_x}$.

The expression $r \frac{\sigma_y}{\sigma_x}$ is known as the *regression coefficient* of y on x ; that is, the deviation of y corresponding on the average to a unit change in the type of x . If we used the other regression line, the slope would be expressed $r \frac{\sigma_x}{\sigma_y}$, which is the regression coefficient of x on y and means that deviation of x which corresponds to a unit change in the type of y . It should be noted that in a perfect correlation, that is, where the variability of the two traits is the same, there are no regression lines and $m = r$.

Let us see what regression coefficients mean in the problem we have solved above, Table XXXVIII.

There it was found that $r = 0.107$ $\sigma_x = 1.408$ $\sigma_y = 3.181$.

$r \frac{\sigma_y}{\sigma_x} = 0.107 \frac{3.181}{1.408} = 0.107 \times 2.359 = 0.241$ regression coefficient of y on x .

$r \frac{\sigma_x}{\sigma_y} = 0.107 \frac{1.408}{3.181} = 0.047$ regression coefficient of x on y .

$y = 0.241x$, which means that for every unit deviation from the type of x (ability in mathematics) it is most prob-

able that there will be an accompanying deviation of 0.241 as much in y (ability in Latin).

$x=0.047y$, which means that for every unit of deviation from the type of y (ability in Latin) it is most probable that there will be an accompanying deviation of $0.047x$ (ability in mathematics).

The correlation coefficient is the geometric mean of the two regression coefficients. That is, the square of the correlation coefficient is equal to the product of the two regression coefficients. Let us represent the regression coefficients by ρ_1 (rho) and ρ_2 respectively. Then $r^2 = \rho_1 \times \rho_2$. This serves as a valuable check on the computation of these coefficients. In the problem cited above $r^2 = 0.114$ and $\rho_1 \times \rho_2 = 0.113$. The difference is due to dropping small fractions in the computation.

The regression coefficients may be taken directly from the values computed in the Ayres Short Method also. In Table XXXIX it was shown that $r = \frac{150}{\sqrt{3,138 \times 616}} = 0.107$.

The fraction $\frac{150}{3138}$ is the regression coefficient of x on y and the fraction $\frac{150}{616}$ is the regression coefficient of y on x . The following computation shows that these fractions are equal to the regression coefficients. Let the regression coefficient

$r \frac{\sigma_x}{\sigma_y}$ be represented by ρ_1 . Then since $r = \frac{\Sigma xy}{N \sigma_x \sigma_y}$,

$$\rho_1 = \frac{\Sigma xy}{N \sigma_x \sigma_y} \times \frac{\sigma_x}{\sigma_y} = \frac{\Sigma xy}{N \sqrt{\frac{\Sigma x^2}{N}} \sqrt{\frac{\Sigma y^2}{N}}} \times \frac{\sqrt{\frac{\Sigma x^2}{N}}}{\sqrt{\frac{\Sigma y^2}{N}}} = \frac{\Sigma xy}{\Sigma y^2} = \frac{150}{3,138}$$

In like manner it may be shown that ρ_2 is equal to the fraction $\frac{150}{616}$.

Coefficients of regressions may be made to yield valuable information in many ways. They may be used in studying such questions as regularity of attendance and

promotion rates; progress through the grades and expenditure for supervision; time spent on spelling drills and scores made in spelling ability; years of schooling and earnings up to a certain age. In fact, coefficients of regressions may be made to yield very valuable information along a great many school lines. In the correlation between Latin and mathematics (cited above) we may use regression coefficients to gain information of the following type.

Suppose a student makes a grade of 85 per cent in mathematics, what is the most probable grade he will make in Latin? From the computation of the correlation and regression coefficients in Table XXXVIII we have the following data:

0.107 = r , the coefficient of correlation,

1.408 = the standard deviation of the mathematics or subject series,

3.181 = the standard deviation of the Latin or relative series;

73.09 = the mean of the Latin series;

80.16 = the mean of the mathematics series.

We also have $y = 0.241x$ which is the equation for the regression of y on x . Since the mean of the subject series is 80.16, a student who gets a grade of 85 in mathematics would be 4.84 above the mean. In Figure XIV we have the means of the two series with the line ST passing through the point of their intersection. The line ST is the line whose equation is $y = 0.241x$. Since this line passes through the origin of the graph (the intersection of the two lines of means) we may assume any values for x and compute the corresponding values for y . If we assume x to be 1, then $y = 0.241$. This means that as the x -values increase one unit above the mean of the subject series, that is, the YOY' axis, the y -values will increase 0.241 as much above the mean of the relative series, that is, the XOX' axis. When, therefore, we assume a value of 85 for x which is 4.84 above

the mean of the subject series, the value of y will be 0.241 as much above the mean of the relative series. That is $0.241 \times 4.84 = 1.166$ which, when added to the mean of the Latin series, $7.309 = 74.26$, which is the most probable grade that would be made in Latin.

The general procedure for predicting ability in one trait when the correlation between two traits and ability in one are given is as follows: Let the trait in which the ability is given be the subject series. Compute the means of both series and write the equation for the regression line using the regression coefficient for y on x . Assume any value for x , as 85 in the problem above. Determine how much this is above or below the x -mean.

Then, in the equation for the regression line, assume x to be 1 and compute the value for y which is 0.241 in the problem above. Multiply the difference between the mean for the x -series and the assumed x -value ($85 - 80.16$) by the value of y and add the product to the y -mean.

It should be noted also that the approximate values may be read directly from the graph of the regression lines. From Figure XIV we note that as we move 5 units to the right of the subject mean the regression line has risen 0.241 as much above the relative mean. We may therefore assume any value for x and read the corresponding values for y directly from the graph.

Assuming values for the mathematics series from 80 to 100, increasing steps of 5, we may accordingly write the corresponding values for y as follows:

THE REGRESSION OF y ON x	
Assumed x -values	Corresponding y -values
80	73.05
85	74.26
90	75.46
95	76.67
100	78.87

If it is desired to assume values for the Latin in the above problem and compute the corresponding mathematics grades, we use the regression of x on y and the equation for the other regression line, which is, $x = 0.047y$.

The line PQ (Fig. XIV) is the line desired. It is measured from the YOY' axis because we desire the regression of x on y . It should be noted that we can assume any value for Latin and read off the approximate mathematics grade directly from the graph, the same as with the other regression line.

The value of the correlation coefficient, r , is always the value of the tangent of the angle that the correlation line makes with the X -axis where it passes through the point of intersection of the X and Y axes. It is never equal to the regression lines. If the correlation is perfect, there is no regression, hence no regression lines. In the above problem $r = 0.107$. Referring to a trigonometric table of natural tangents we find that 0.107 is the tangent of the angle $6^\circ 7'$.

Therefore, if we desired to draw the correlation line in the above correlation Table XXXVIII, it would make an angle of $6^\circ 7'$ with the X -axis.

The significance of the tangent in determining the slope of the correlation line and also the reasons why the value of r , the coefficient of correlation, may vary from -1 through 0 to $+1$ may be made clearer by geometry in the following illustration.

About O as a center, describe a circle with a radius equal to 1.

Let XOX' be a diameter of the circle and YOY' another diameter perpendicular to it. Let CX' be a geometrical tangent to the circle O drawn perpendicular to the diameter XOX' , and let OC be a line bisecting the angle YOX' and intersecting the tangent CX' at C . Then, by geometry, we know that the triangle COX' is a right angle triangle, that the angles COX' and $X'CO$ are each 45° , and that the side

CX' = the side OX' . By construction $OX' = 1$. The line CX' therefore, equals 1, and since the line CX' is tangent to the circle at X' , the numerical value of such a tangent, limited by the point of intersection of a line forming an angle of 45° with the horizontal axis, is 1.

From O draw the line OC' , making the angle $C'OX'$ equal to $6^\circ 7'$. Now since $CX' = OX' = 1$, and since the value of

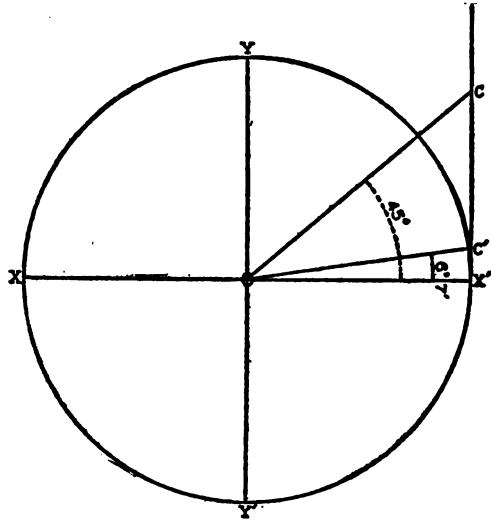


FIGURE XIII

the coefficient of correlation, is 0.107, then the value of $C'X'$ as a percentage of 1 is $\frac{C'X'}{CX'} = \frac{10.7}{100}$ or 0.107. Now it is evident that if the angle $C'OX'$ were zero, the tangent would be zero and hence r would equal zero. It is also evident that r can never be greater than 1 because when the angle becomes greater than 45° we measure the tangent from the angle made by the correlation line and the line YOY . If

the correlation line should swing to the left of YOY' into the second quadrant, then r would be negative because the x -values would grow less as the y -values increased. The tangent would reach its maximum length when the correlation line reached 135° from point X' on the circle and would have a value of -1 . Hence the values of r will vary from $+1$ through 0 to -1 .

Figure XIV shows the correlation line and the two regression lines for the data in Table XXXVIII. In that table the mean for the mathematics series is 80.16 and is represented in Figure XHI by the line YOY' . The mean for the Latin series is 73.09 and is represented by the line XOX' . The line MN is the correlation line and is drawn so as to make an angle of $6^\circ 7'$ with the x -axis. The line CD is the line of perfect correlation and makes an angle of 45° with the X -axis. The two regression lines are ST , the regression of y on x , which makes an angle of $13^\circ 33'$ with the X -axis, and PQ , the regression of x on y which makes an angle of $2^\circ 42'$ with the Y -axis.

The Reliability of the Correlation Coefficient.—We found the coefficient of correlation between abilities in Latin and in mathematics to be 0.107. The number of cases taken to get this value was 310. This number represents but a small part of the people in high school taking Latin and mathematics. The question arises: If we were to take 310 other cases and compute the correlation coefficient, would it be 0.107? Or, if we took all the students taking Latin and mathematics in high schools and computed the correlation between the two abilities, would it still be 0.107? We cannot answer these questions dogmatically and say the coefficient of correlation thus computed would, or would not, be 0.107, but we can apply the laws of "chance" and speak rather dogmatically from the facts gained. We know that the reliability of the coefficient of correlation, just as the reliability of the mean or of the standard devia-

tion, depends on the *normality* of the distributions in question. If the distributions are approximately normal, the coefficient of correlation will be fairly reliable. On the

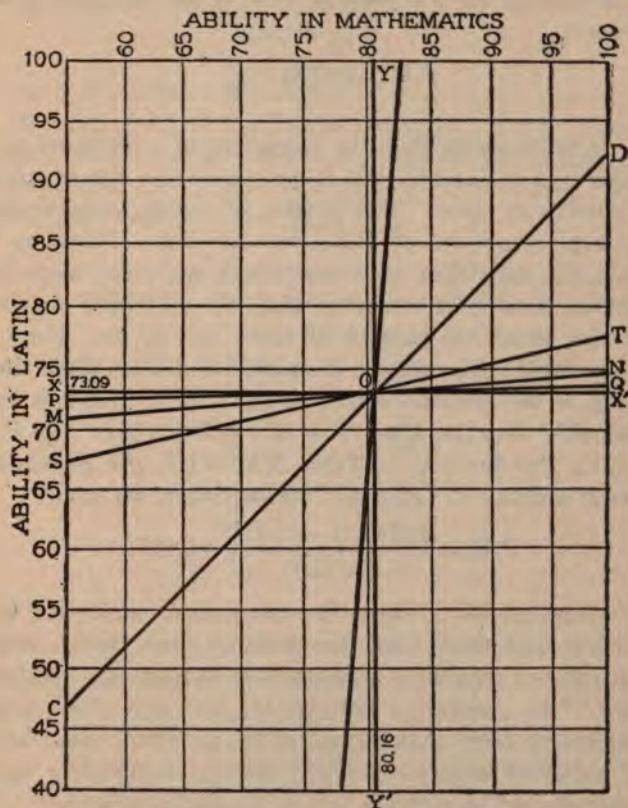


FIGURE XIV. SHOWING THE POSITION OF THE CORRELATION LINE AND REGRESSION LINES FOR DATA IN TABLE XXXVIII

other hand, if the distributions do not approach normality, the reliability of the coefficient of correlation becomes less.

When the distributions resemble normality, then the

probable error (P.E.) may be used to estimate the *probable stability* of the coefficient. We find from the normal probability curve that $P.E. = 0.6745 \sigma$.

The formula for the *probable error of the coefficient of correlation* is

$$P.E._r = 0.6745 \frac{1-r^2}{\sqrt{N}}$$

This formula shows that the reliability of r increases as N increases; not directly, but in proportion to the square of the number of cases. This is true, of course, because when N is large, the value of $P.E._r$ becomes less. Therefore, to double the reliability of a coefficient we must take four times the number of cases; to triple the reliability we must take nine times the number of cases, and so on. For r to be considered fairly reliable it should be at least three times as large as the probable error, on the ground that it is very improbable that the true value of r falls outside $r \pm 3 P.E.$ Applying this formula to Table XXXVIII, the correlation between abilities in Latin and mathematics, we have

$$P.E._r = \frac{0.6745 (1 - 0.107^2)}{\sqrt{310}} = 0.037$$

or $r = 0.107 \pm 0.037$. Here the correlation coefficient is a little less than three times the probable error, which makes it not entirely reliable as a measure of correlation. Whipple says:⁹ "In general, a correlation, like any other determination, to have claim to scientific attention must be at least twice as large as its P.E., and to be perfectly satisfactory, should be perhaps four or five times as large."

Spearman's Method of Rank Correlation.—Spearman's "foot-rule" for measuring correlation is a simple method of comparison by "rank" or "position" rather in terms of absolute quantity. This method is becoming popular

⁹ *Manual of Mental and Physical Tests*, Part I, p. 40.

because of the ease of computation. It, of course, is less accurate than the product-moment method by Pearson. The formula for this method is

$$R = 1 - \frac{6\Sigma g}{N^2 - 1}$$

in which R expresses the degree of rank correlation (not to be confused with r in the Pearson formula); g is the numerical gains in rank of an individual in the second, as compared with the first series; and N is the number of cases. Table XLI illustrates the computation of R .

TABLE XLI.—CORRELATION BETWEEN A CLASS IN ADDITION AND HANDWRITING MEASURED BY THE THORNDIKE SCALE

Pupil	Addition	Hand-writing	Gains, g
A.....	13.....	16.....	0
B.....	11.....	14.....	0
C.....	10.....	15.....	1
D.....	9.....	11.....	0
E.....	8.....	13.....	1
F.....	7.....	9.....	0
G.....	6.....	12.....	2
			<hr/> 4

Substituting in the formula $R = 1 - \frac{6\Sigma g}{N^2 - 1} = 1 - \frac{24}{48} = 0.5$.

In using the data in Table XLI to illustrate the computation of the degree of correlation by the Spearman Rank Method we proceed as follows:

Rank the measures in each series in the order of their magnitude. We may start with either the largest or smallest scores in ranking them. If we start with the largest scores in one series, we must do the same in the other, and vice versa. Compute the amount of gain in rank of each score value in the second series over the rank of its cor-

responding score in the first series. For example: pupils A, B, D, and F did not gain in rank in handwriting over their respective ranks in addition. Pupil c ranks third in addition but second in handwriting, hence his gain in rank is 1. Pupil E ranks fifth in addition and fourth in handwriting, hence gains 1. Pupil G ranks seventh in addition and fifth in handwriting, hence gains 2. The sum of the gains is 4 which when substituted in the formula gives a value for R equal to 0.5.

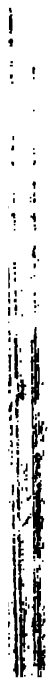
In case of a tie in the rank in either series it is customary to divide the ranks in such a manner as to keep the total number of ranks in each series the same. If, for example, two scores ranked fifth each should be assigned a value of 5.5 (that is one-half of $5+6$). If three ranked fifth, they should all be assigned the rank of 6 (the mean of the fifth, sixth, and seventh places).

The rank method should be used only when N is small, in which case its reliability is about as great as the more accurate product-moment method.

BIBLIOGRAPHY

1. AYRES, LEONARD P., "A Shorter Method of Computing the Coefficient of Correlation," *Journal of Educational Research*, Vol. 1, March, 1920; also "The Application of Tables of Distribution of a Shorter Method for Computing Coefficients of Correlation," *Journal of Educational Research*, Vol. 1, April, 1920.
2. BOWLEY, ARTHUR L., *An Elementary Manual of Statistics* (MacDonald and Evans, London, 1910).
3. BOWLEY, ARTHUR L., *The Nature and Purpose of the Measurement of Social Phenomena* (P. S. King & Son, Ltd., London, 1915).
4. DAVENPORT, EUGENE, *Principles of Breeding* (Ginn & Co., 1907).
5. ELDETON, W. PALIN, and ELDETON, ETHEL M., *Primer of Statistics* (Adam and Charles Black, London, 1914).
6. JUDD, CHARLES H., *Introduction to the Scientific Study of Education* (Ginn & Co., 1918).

7. KING, WILLFORD I., *The Elements of Statistical Method* (The Macmillan Co., 1912).
8. MCCALL, WILLIAM A., "How to Compute a Median," *Teachers College Record*, Vol. 21, March, 1920.
9. MCCALL, WILLIAM A., *How to Measure in Education* (The Macmillan Co., 1922).
10. MONROE, WALTER S., *Measuring the Results of Teaching* (Houghton Mifflin Co., 1918).
11. PEARSON, KARL, *The Grammar of Science* (Adam and Charles Black, London, 1911).
12. ROBERTS, HERBERT F., "A Practical Method of Demonstrating the Error of Mean Square," *School Science and Mathematics*, Vol. 19, pp. 667-692.
13. ROBERTS, HERBERT F., "A Demonstration of the Coefficient of Correlation for Elementary Students in Plant Breeding," *School Science and Mathematics*, Vol. 19, pp. 619-628.
14. RUGG, HAROLD O., *Statistical Methods Applied to Education* (Houghton Mifflin Co., 1917).
15. SECRIST, HORACE, *An Introduction to Statistical Methods*.
16. STARCH, DANIEL, *Educational Measurements* (The Macmillan Co., 1917).
17. National Society for the Study of Education, *Twenty-first Yearbook*, Part II (Public School Publishing Co., Bloomington, Ill., 1922).
18. THEISEN, W., *Report on the Use of Some Standard Tests for 1916-1917* (Wisconsin State Department of Public Instruction, Madison, Wis., 1918).
19. THORNDIKE, EDWARD L., *An Introduction to the Theory of Mental and Social Measurements* (Teachers College, Columbia University, New York, 1916).
20. WEST, CARL J., *Introduction to Mathematical Statistics* (R. G. Adams & Co., Columbus, O., 1918).
21. WHIPPLE, GUY M., *Manual of Physical and Mental Tests*, Part I (Warwick and York, 1920).
22. WILSON, G. M., and HOKE, KREMER J., *How to Measure* (The Macmillan Co., 1920).
23. Various Conferences on Educational Measurement; Indiana University Bulletins (University of Indiana, Bloomington, Ind.)



APPENDIX

SQUARES AND SQUARE ROOTS

No.	Square	Square Root	No.	Square	Square Root	No.	Square	Square Root
1	1	1.000	36	12 96	6.000	71	50 41	8.426
2	4	1.414	37	13 69	6.083	72	51 84	8.485
3	9	1.732	38	14 44	6.164	73	53 29	8.544
4	16	2.000	39	15 21	6.245	74	54 76	8.602
5	25	2.236	40	16 00	6.325	75	56 25	8.660
6	36	2.449	41	16 81	6.403	76	57 76	8.718
7	49	2.646	42	17 64	6.481	77	59 29	8.775
8	64	2.828	43	18 49	6.557	78	60 84	8.832
9	81	3.000	44	19 36	6.633	79	62 41	8.888
10	1 00	3.162	45	20 25	6.708	80	64 00	8.944
11	1 21	3.317	46	21 16	6.782	81	65 61	9.000
12	1 44	3.464	47	22 09	6.856	82	67 24	9.055
13	1 69	3.606	48	23 04	6.928	83	68 89	9.110
14	1 96	3.742	49	24 01	7.000	84	70 56	9.165
15	2 25	3.873	50	25 00	7.071	85	72 25	9.220
16	2 56	4.000	51	26 01	7.141	86	73 96	9.274
17	2 89	4.123	52	27 04	7.211	87	75 69	9.327
18	3 24	4.243	53	28 09	7.280	88	77 44	9.381
19	3 61	4.359	54	29 16	7.348	89	79 21	9.434
20	4 00	4.472	55	30 25	7.416	90	81 00	9.487
21	4 41	4.583	56	31 36	7.483	91	82 81	9.539
22	4 84	4.690	57	32 49	7.550	92	84 64	9.592
23	5 29	4.796	58	33 64	7.616	93	86 49	9.644
24	5 76	4.899	59	34 81	7.681	94	88 36	9.695
25	6 25	5.000	60	36 00	7.746	95	90 25	9.747
26	6 76	5.099	61	37 21	7.810	96	92 16	9.798
27	7 29	5.196	62	38 44	7.874	97	94 09	9.849
28	7 84	5.292	63	39 69	7.937	98	96 04	9.899
29	8 41	5.385	64	40 96	8.000	99	98 01	9.950
30	9 00	5.477	65	42 25	8.062	100	1 00 00	10.000
31	9 61	5.568	66	43 56	8.124	101	1 02 01	10.050
32	10 24	5.657	67	44 89	8.185	102	1 04 04	10.100
33	10 89	5.745	68	46 24	8.246	103	1 06 09	10.149
34	11 56	5.831	69	47 61	8.307	104	1 08 16	10.198
35	12 25	5.916	70	49 00	8.367	105	1 10 25	10.247

SQUARES AND SQUARE ROOTS—Continued

No.	Square	Square Root	No.	Square	Square Root	No.	Square	Square Root
106	1 12 36	10.296	141	1 98 81	11.874	176	2 09 76	13.266
107	1 14 49	10.344	142	2 01 64	11.916	177	3 13 29	13.304
108	1 16 64	10.392	143	2 04 49	11.958	178	3 16 84	13.342
109	1 18 81	10.440	144	2 07 36	12.000	179	3 20 41	13.379
110	1 21 00	10.488	145	2 10 25	12.042	180	3 24 00	13.416
111	1 23 21	10.536	146	2 13 16	12.083	181	3 27 61	13.454
112	1 25 44	10.583	147	2 16 09	12.124	182	2 31 24	13.491
113	1 27 69	10.630	148	2 19 04	12.166	183	3 34 89	13.528
114	1 29 96	10.677	149	2 22 01	12.207	184	3 38 56	13.565
115	1 32 25	10.724	150	2 25 00	12.247	185	3 42 25	13.601
116	1 34 56	10.770	151	2 28 01	12.288	186	3 45 96	13.638
117	1 36 89	10.817	152	2 31 04	12.329	187	3 49 69	13.675
118	1 39 24	10.863	153	2 34 09	12.369	188	3 53 44	13.711
119	1 41 61	10.909	154	2 37 16	12.410	189	3 57 21	13.748
120	1 44 00	10.954	155	2 40 25	12.450	190	3 61 00	13.784
121	1 46 41	11.000	156	2 43 36	12.490	191	3 64 81	13.820
122	1 48 84	11.045	157	2 46 49	12.530	192	3 68 64	13.856
123	1 51 29	11.091	158	2 49 64	12.570	193	3 72 49	13.892
124	1 53 76	11.136	159	2 52 81	12.610	194	3 76 36	13.928
125	1 56 25	11.180	160	2 56 00	12.649	195	3 80 25	13.964
126	1 58 76	11.225	161	2 59 21	12.689	196	3 84 16	14.000
127	1 61 29	11.269	162	2 62 44	12.728	197	3 88 09	14.036
128	1 63 84	11.314	163	2 65 69	12.767	198	3 92 04	14.071
129	1 66 41	11.358	164	2 68 96	12.806	199	3 96 01	14.107
130	1 69 00	11.402	165	2 72 25	12.845	200	4 00 00	14.142
131	1 71 61	11.446	166	2 75 56	12.884	201	4 04 01	14.177
132	1 74 24	11.489	167	2 78 89	12.923	202	4 08 04	14.213
133	1 76 89	11.533	168	2 82 24	12.961	203	4 12 09	14.248
134	1 79 56	11.576	169	2 85 61	13.000	204	4 16 16	14.283
135	1 82 25	11.619	170	2 89 00	13.038	205	4 20 25	14.318
136	1 84 96	11.662	171	2 92 41	13.077	206	4 24 36	14.353
137	1 87 69	11.705	172	2 95 84	13.115	207	4 28 49	14.387
138	1 90 44	11.747	173	2 99 29	13.153	208	4 32 64	14.422
139	1 93 21	11.790	174	3 02 76	13.191	209	4 36 81	14.457
140	1 96 00	11.832	175	3 06 25	13.229	210	4 41 00	14.491

SQUARES AND SQUARE ROOTS

369

SQUARES AND SQUARE ROOTS—Continued

No.	Square	Square Root	No.	Square	Square Root	No.	Square	Square Root
211	4 45 21	14.526	246	6 05 16	15.684	281	7 89 61	16.763
212	4 49 44	14.560	247	6 10 09	15.716	282	7 95 24	16.793
213	4 53 09	14.595	248	6 15 04	15.748	283	8 00 89	16.823
214	4 57 96	14.629	249	6 20 01	15.780	284	8 06 56	16.852
215	4 62 25	14.663	250	6 25 00	15.811	285	8 12 25	16.882
216	4 66 56	14.697	251	6 30 01	15.843	286	8 17 96	16.912
217	4 70 89	14.731	252	6 35 04	15.875	287	8 23 69	16.941
218	4 75 24	14.765	253	6 40 09	15.906	288	8 29 44	16.971
219	4 79 61	14.799	254	6 45 16	15.937	289	8 35 21	17.000
220	4 84 00	14.832	255	6 50 25	15.969	290	8 41 00	17.029
221	4 88 41	14.866	256	6 55 36	16.000	291	8 46 81	17.059
222	4 92 84	14.900	257	6 60 49	16.031	292	8 52 64	17.088
223	4 97 29	14.933	258	6 65 64	16.062	293	8 58 49	17.117
224	5 01 76	14.967	259	6 70 81	16.093	294	8 64 36	17.146
225	5 06 25	15.000	260	6 76 00	16.125	295	8 70 25	17.176
226	5 10 76	15.033	261	6 81 21	16.155	296	8 76 16	17.205
227	5 15 29	15.067	262	6 86 44	16.186	297	8 82 09	17.234
228	5 19 84	15.100	263	6 91 69	16.217	298	8 88 04	17.263
229	5 24 41	15.133	264	6 96 96	16.248	299	8 94 01	17.292
230	5 29 00	15.166	265	7 02 25	16.279	300	9 00 00	17.321
231	5 33 61	15.199	266	7 07 56	16.310	301	9 06 01	17.349
232	5 38 24	15.232	267	7 12 89	16.340	302	9 12 04	17.378
233	5 42 89	15.264	268	7 18 24	16.371	303	9 18 09	17.407
234	5 47 56	15.297	269	7 23 61	16.401	304	9 24 16	17.436
235	5 52 25	15.330	270	7 29 00	16.432	305	9 30 25	17.464
236	5 56 96	15.362	271	7 34 41	16.462	306	9 36 36	17.493
237	5 61 69	15.395	272	7 39 84	16.492	307	9 42 49	17.521
238	5 66 44	15.427	273	7 45 29	16.523	308	9 48 64	17.550
239	5 71 21	15.460	274	7 50 76	16.553	309	9 54 81	17.578
240	5 76 00	15.492	275	7 56 25	16.583	310	9 61 00	17.607
241	5 80 81	15.524	276	7 61 76	16.613	311	9 67 21	17.635
242	5 85 64	15.556	277	7 67 29	16.643	312	9 73 44	17.664
243	5 90 49	15.588	278	7 72 84	16.673	313	9 79 69	17.692
244	5 95 36	15.620	279	7 78 41	16.703	314	9 85 96	17.720
245	6 00 25	15.652	280	7 84 00	16.733	315	9 92 25	17.748

SQUARES AND SQUARE ROOTS—Continued

No.	Square	Square Root	No.	Square	Square Root	No.	Square	Square Root
316	9 98 56	17.776	351	12 32 01	18.735	386	14 89 96	19.647
317	10 04 89	17.804	352	12 39 04	18.762	387	14 97 69	19.672
318	10 11 24	17.833	353	12 46 09	18.788	388	15 05 44	19.698
319	10 17 61	17.861	354	12 53 16	18.815	389	15 13 21	19.723
320	10 24 00	17.889	355	12 60 25	18.841	390	15 21 00	19.748
321	10 30 41	17.916	356	12 67 36	18.868	391	15 28 81	19.774
322	10 36 84	17.944	357	12 74 49	18.894	392	15 36 64	19.799
323	10 43 29	17.972	358	12 81 64	18.921	393	15 44 49	19.824
324	10 49 76	18.000	359	12 88 81	18.947	394	15 52 36	19.849
325	10 56 25	18.028	360	12 96 00	18.974	395	15 60 25	19.875
326	10 62 76	18.055	361	13 03 21	19.000	396	15 68 16	19.900
327	10 69 29	18.083	362	13 10 44	19.026	397	15 76 09	19.925
328	10 75 84	18.111	363	13 17 69	19.053	398	15 84 04	19.950
329	10 82 41	18.138	364	13 24 96	19.079	399	15 92 01	19.975
330	10 89 00	18.166	365	13 32 25	19.105	400	16 00 00	20.000
331	10 95 61	18.193	366	13 39 56	19.131	401	16 08 01	20.025
332	11 02 24	18.221	367	13 46 89	19.157	402	16 16 04	20.050
333	11 08 89	18.248	368	13 54 24	19.183	403	16 24 09	20.075
334	11 15 56	18.276	369	13 61 61	19.209	404	16 32 16	20.100
335	11 22 25	18.303	370	13 69 00	19.235	405	16 40 25	20.125
336	11 28 96	18.330	371	13 76 41	19.261	406	16 48 36	20.149
337	11 35 69	18.358	372	13 83 84	19.287	407	16 56 49	20.174
338	11 42 44	18.385	373	13 91 29	19.313	408	16 64 64	20.199
339	11 49 21	18.412	374	13 98 76	19.339	409	16 72 81	20.224
340	11 56 00	18.439	375	14 06 25	19.365	410	16 81 00	20.248
341	11 62 81	18.466	376	14 13 76	19.391	411	16 89 21	20.273
342	11 69 64	18.493	377	14 21 29	19.416	412	16 97 44	20.298
343	11 76 49	18.520	378	14 28 84	19.442	413	17 05 69	20.322
344	11 83 36	18.547	379	14 36 41	19.468	414	17 13 96	20.347
345	11 90 25	18.574	380	14 44 00	19.494	415	17 22 25	20.372
346	11 97 16	18.601	381	14 51 61	19.519	416	17 30 56	20.396
347	12 04 09	18.628	382	14 59 24	19.545	417	17 38 89	20.421
348	12 11 04	18.655	383	14 66 89	19.570	418	17 47 24	20.445
349	12 18 01	18.682	384	14 74 56	19.596	419	17 55 61	20.469
350	12 25 00	18.708	385	14 82 25	19.621	420	17 64 00	20.494

SQUARES AND SQUARE ROOTS—Continued

No.	Square	Square Root	No.	Square	Square Root	No.	Square	Square Root
421	17 72 41	20.518	456	20 79 36	21.354	491	24 10 81	22.159
422	17 80 84	20.543	457	20 88 49	21.378	492	24 20 64	22.181
423	17 89 29	20.567	458	21 06 81	21.401	493	24 30 49	22.204
424	17 97 76	20.591	459	21 16 00	21.424	494	24 40 36	22.226
425	18 06 25	20.616	460	21 16 00	21.448	495	24 50 25	22.249
426	18 14 76	20.640	461	21 25 21	21.471	496	24 60 16	22.271
427	18 23 29	20.664	462	21 34 44	21.494	497	24 70 09	22.293
428	18 31 84	20.688	463	21 43 69	21.517	498	24 80 04	22.316
429	18 40 41	20.712	464	21 52 96	21.541	499	24 90 01	22.338
430	18 49 00	20.736	465	21 62 25	21.564	500	25 00 00	22.361
431	18 57 61	20.761	466	21 71 56	21.587	501	25 10 01	22.383
432	18 66 24	20.785	467	21 80 89	21.610	502	25 20 04	22.405
433	18 74 89	20.809	468	21 90 24	21.633	503	25 30 09	22.428
434	18 83 56	20.833	469	21 99 61	21.656	504	25 40 16	22.450
435	18 92 25	20.857	470	22 09 00	21.679	505	25 50 25	22.472
436	19 00 96	20.881	471	22 18 41	21.703	506	25 60 36	22.494
437	19 09 69	20.905	472	22 27 84	21.726	507	25 70 49	22.517
438	19 18 44	20.928	473	22 37 29	21.749	508	25 80 64	22.539
439	19 27 21	20.952	474	22 46 76	21.772	509	25 90 81	22.561
440	19 36 00	20.976	475	22 56 25	21.794	510	26 01 00	22.583
441	19 44 81	21.000	476	22 65 76	21.817	511	26 11 21	22.605
442	19 53 64	21.024	477	22 75 29	21.840	512	26 21 44	22.627
443	19 62 49	21.048	478	22 84 84	21.863	513	26 31 69	22.650
444	19 71 36	21.071	479	22 94 41	21.886	514	26 41 96	22.672
445	19 80 25	21.095	480	23 04 00	21.909	515	26 52 25	22.694
446	19 89 16	21.119	481	23 13 61	21.932	516	26 62 56	22.716
447	19 98 09	21.142	482	23 23 24	21.954	517	26 72 89	22.738
448	20 07 04	21.166	483	23 32 89	21.977	518	26 83 24	22.760
449	20 16 01	21.190	484	23 42 56	22.000	519	26 93 61	22.782
450	20 25 00	21.213	485	23 52 25	22.023	520	27 04 00	22.804
451	20 34 01	21.237	486	23 61 96	22.045	521	27 14 41	22.825
452	20 43 04	21.260	487	23 71 69	22.068	522	27 24 84	22.847
453	20 52 09	21.284	488	23 81 44	22.091	523	27 35 29	22.869
454	20 61 16	21.307	489	23 91 21	22.113	524	27 45 76	22.891
455	20 70 25	21.331	490	24 01 00	22.136	525	27 56 25	22.913

SQUARES AND SQUARE ROOTS—Continued

No.	Square	Square Root	No.	Square	Square Root	No.	Square	Square Root
526	27 66 76	22.935	561	31 47 21	23.685	596	35 52 16	24.413
527	27 77 29	22.956	562	31 58 44	23.707	597	35 64 09	24.434
528	27 87 84	22.978	563	31 69 69	23.728	598	35 76 04	24.454
529	27 98 41	23.000	564	31 80 96	23.749	599	35 88 01	24.474
530	28 09 00	23.022	565	31 92 25	23.770	600	36 00 00	24.495
531	28 19 61	23.043	566	32 03 56	23.791	601	36 12 01	24.515
532	28 30 24	23.065	567	32 14 89	23.812	602	36 24 04	24.536
533	28 40 89	23.087	568	32 26 24	23.833	603	36 36 09	24.556
534	28 51 56	23.108	569	32 37 61	23.854	604	36 48 16	24.576
535	28 62 25	23.130	570	32 49 00	23.875	605	36 60 25	24.597
536	28 72 96	23.152	571	32 60 41	23.896	606	36 72 36	24.617
537	28 83 69	23.173	572	32 71 84	23.917	607	36 84 49	24.637
538	28 94 44	23.195	573	32 83 29	23.937	608	36 96 64	24.658
539	29 05 21	23.216	574	32 94 76	23.958	609	37 08 81	24.678
540	29 16 00	23.238	575	33 06 25	23.979	610	37 21 00	24.698
541	29 26 81	23.259	576	33 17 76	24.000	611	37 33 21	24.718
542	29 37 64	23.281	577	33 29 29	24.021	612	37 45 44	24.739
543	29 48 49	23.302	578	33 40 84	24.042	613	37 57 69	24.759
544	29 59 36	23.324	579	33 52 41	24.062	614	37 69 96	24.779
545	29 70 25	23.345	580	33 64 00	24.083	615	37 82 25	24.799
546	29 81 16	23.367	581	33 75 61	24.104	616	37 94 56	24.819
547	29 92 09	23.388	582	33 87 24	24.125	617	38 06 89	24.839
548	30 03 04	23.409	583	33 98 89	24.145	618	38 19 24	24.860
549	30 14 01	23.431	584	34 10 56	24.166	619	38 31 61	24.880
550	30 25 00	25.452	585	34 22 25	24.187	620	38 44 00	24.900
551	30 36 01	23.473	586	34 33 96	24.207	621	38 56 41	24.920
552	30 47 04	23.495	587	34 45 69	24.228	622	38 68 84	24.940
553	30 58 09	23.516	588	34 57 44	24.249	623	38 81 29	24.960
554	30 69 16	23.537	589	34 69 21	24.269	624	38 93 76	24.980
555	30 80 25	23.558	590	34 81 00	24.290	625	39 06 25	25.000
556	30 91 36	23.580	591	34 92 81	24.310	626	39 18 76	25.020
557	31 02 49	23.601	592	35 04 64	24.331	627	39 31 29	25.040
558	31 13 64	23.622	593	35 16 49	24.352	628	39 43 84	25.060
559	31 24 81	23.643	594	35 28 36	24.372	629	39 56 41	25.080
560	31 36 00	23.664	595	35 40 25	24.393	630	39 69 00	25.100

SQUARES AND SQUARE ROOTS

373

SQUARES AND SQUARE ROOTS—Continued

No.	Square	Square Root	No.	Square	Square Root	No.	Square	Square Root
631	39 81 61	25.120	666	44 35 56	25.807	701	49 14 01	26.476
632	39 94 24	25.140	667	44 48 89	25.826	702	49 28 04	26.495
633	40 06 89	25.159	668	44 62 24	25.846	703	49 42 09	26.514
634	40 19 56	25.179	669	44 75 61	25.865	704	49 56 16	26.533
635	40 32 25	25.199	670	44 89 00	25.884	705	49 70 25	26.552
636	40 44 96	25.219	671	45 02 41	25.904	706	49 84 36	26.571
637	40 57 69	25.239	672	45 15 84	25.923	707	49 98 49	26.589
638	40 70 44	25.259	673	45 29 29	25.942	708	50 12 64	26.608
639	40 83 21	25.278	674	45 42 76	25.962	709	50 26 81	26.627
640	40 96 00	25.298	675	45 56 25	25.981	710	50 41 00	26.646
641	41 08 81	25.318	676	45 69 76	26.000	711	50 55 21	26.665
642	41 21 64	25.338	677	45 83 29	26.019	712	50 69 44	26.683
643	41 34 49	25.357	678	45 96 84	26.038	713	50 83 69	26.702
644	41 47 36	25.377	679	46 10 41	26.058	714	50 97 96	26.721
645	41 60 25	25.397	680	46 24 00	26.077	715	51 12 25	26.739
646	41 73 16	25.417	681	46 37 61	26.096	716	51 26 56	26.758
647	41 86 09	25.436	682	46 51 24	26.115	717	51 40 89	26.777
648	41 99 04	25.456	683	46 64 89	26.134	718	51 55 24	26.796
649	42 12 01	25.475	684	46 78 56	26.153	719	51 69 61	26.814
650	42 26 00	25.495	685	46 92 25	26.173	720	51 84 00	26.833
651	42 38 01	25.515	686	47 05 96	26.192	721	51 98 41	26.851
652	42 51 04	25.534	687	47 19 69	26.211	722	52 12 84	26.870
653	42 64 09	25.554	688	47 33 44	26.230	723	52 27 29	26.889
654	42 77 16	25.593	689	47 47 21	26.249	724	52 41 76	26.907
655	42 90 25	25.573	690	47 61 00	26.268	725	52 56 25	26.926
656	43 03 36	25.612	691	47 74 81	26.287	726	52 70 76	26.944
657	43 16 49	25.632	692	47 88 64	26.306	727	52 85 29	26.963
658	43 29 64	25.652	693	48 02 49	26.325	728	52 99 84	26.981
659	43 42 81	25.671	694	48 16 36	26.344	729	53 14 41	27.000
660	43 56 00	25.690	695	48 30 25	26.363	730	53 29 00	27.019
661	43 69 21	25.710	696	48 44 16	26.382	731	53 43 61	27.037
662	43 82 44	25.729	697	48 58 09	26.401	732	53 58 24	27.055
663	43 95 69	25.749	698	48 72 04	26.420	733	53 72 89	27.074
664	44 08 96	25.768	699	48 86 01	26.439	734	53 87 56	27.092
665	44 22 25	25.788	700	49 00 00	26.458	735	54 02 25	27.111

SQUARES AND SQUARE ROOTS—Continued

No.	Square	Square Root	No.	Square	Square Root	No.	Square	Square Root
736	54 16 96	27.129	771	59 44 41	27.767	806	64 96 36	28.390
737	54 31 69	27.148	772	59 59 84	27.785	807	65 12 49	28.408
738	54 46 44	27.166	773	59 75 29	27.803	808	65 28 64	28.425
739	54 61 21	27.185	774	59 90 76	27.821	809	65 44 81	28.443
740	54 76 00	27.203	775	60 06 25	27.839	810	65 61 00	28.460
741	54 90 81	27.221	776	60 21 76	27.857	811	65 77 21	28.478
742	55 05 64	27.240	777	60 37 29	27.875	812	65 93 44	28.496
743	55 20 49	27.258	778	60 52 84	27.893	813	66 09 69	28.513
744	55 35 36	27.276	779	60 68 41	27.911	814	66 25 96	28.531
745	55 50 25	27.295	780	60 84 00	27.928	815	66 42 25	28.548
746	55 65 16	27.313	781	60 99 61	27.946	816	66 58 56	28.566
747	55 80 09	27.331	782	61 15 24	27.964	817	66 74 89	28.583
748	55 95 04	27.350	783	61 30 89	27.982	818	66 91 24	28.601
749	56 10 01	27.368	784	61 46 56	28.000	819	67 07 61	28.618
750	56 25 00	27.386	785	61 62 25	28.018	820	67 24 00	28.636
751	56 40 01	27.404	786	61 77 96	28.036	821	67 40 41	28.653
752	56 55 04	27.423	787	61 93 69	28.054	822	67 56 84	28.671
753	56 70 09	27.441	788	62 09 44	28.071	823	67 73 29	28.688
754	56 85 16	27.459	789	62 25 21	28.089	824	67 89 76	28.705
755	57 00 25	27.477	790	62 41 00	28.107	825	68 06 25	28.723
756	57 15 36	27.495	791	62 56 81	28.125	826	68 22 76	28.740
757	57 30 49	27.514	792	62 72 64	28.142	827	68 39 29	28.758
758	57 45 64	27.532	793	62 88 49	28.160	828	68 55 84	28.775
759	57 60 81	27.550	794	63 04 36	28.178	829	68 72 41	28.792
760	57 76 00	27.568	795	63 20 25	28.196	830	68 89 00	28.810
761	57 91 21	27.586	796	63 36 16	28.213	831	69 05 61	28.827
762	58 06 44	27.604	797	63 52 09	28.231	832	69 22 24	28.844
763	58 21 69	27.622	798	63 68 04	28.249	833	69 38 89	28.862
764	58 36 96	27.641	799	63 84 01	28.267	834	69 55 56	28.879
765	58 52 25	27.659	800	64 00 00	28.284	835	69 72 25	28.896
766	58 67 56	27.677	801	64 16 01	28.302	836	69 88 96	28.914
767	58 82 89	27.695	802	64 32 04	28.320	837	70 05 69	28.931
768	58 98 24	27.713	803	64 48 09	28.337	838	70 22 44	28.948
769	59 13 61	27.731	804	64 64 16	28.355	839	70 39 21	28.965
770	59 29 00	27.749	805	64 80 25	28.373	840	70 56 00	28.983

SQUARES AND SQUARE ROOTS

375

SQUARES AND SQUARE ROOTS—Continued

No.	Square	Square Root	No.	Square	Square Root	No.	Square	Square Root
841	70 72 81	29.000	876	76 73 76	29.597	911	82 99 21	30.183
842	70 89 64	29.017	877	76 91 29	29.614	912	83 17 44	30.199
843	71 06 49	29.034	878	77 08 84	29.631	913	83 35 69	30.216
844	71 23 36	29.052	879	77 26 41	29.648	914	83 53 96	30.232
845	71 40 25	29.069	880	77 44 00	29.665	915	83 72 25	30.249
846	71 57 16	29.086	881	77 61 61	29.682	916	83 90 56	30.265
847	71 74 09	29.103	882	77 79 24	29.698	917	84 08 89	30.282
848	71 91 04	29.120	883	77 96 89	29.715	918	84 27 24	30.299
849	72 08 01	29.138	884	78 14 56	29.732	919	84 45 61	30.315
850	72 25 00	29.155	885	78 32 25	29.749	920	84 64 00	30.332
851	72 42 01	29.172	886	78 49 96	29.766	921	84 82 41	30.348
852	72 59 04	29.189	887	78 67 69	29.783	922	85 00 84	30.364
853	72 76 09	29.206	888	78 85 44	29.799	923	85 19 29	30.381
854	72 93 16	29.223	889	79 03 21	29.816	924	85 37 76	30.397
855	73 10 25	29.240	890	79 21 00	29.833	925	85 56 25	30.414
856	73 27 36	29.257	891	79 38 81	29.850	926	85 74 76	30.430
857	73 44 49	29.275	892	79 56 64	29.866	927	85 93 29	30.447
858	73 61 64	29.292	893	79 74 49	29.883	928	86 11 84	30.463
859	73 78 81	29.309	894	79 92 36	29.900	929	86 30 41	30.480
860	73 96 00	29.326	895	80 10 25	29.916	930	86 49 00	30.496
861	74 13 21	29.343	896	80 28 16	29.933	931	86 67 61	30.512
862	74 30 44	29.360	897	80 46 09	29.950	932	86 86 24	30.529
863	74 47 69	29.377	898	80 64 04	29.967	933	87 04 89	30.545
864	74 64 96	29.394	899	80 82 01	29.983	934	87 23 56	30.561
865	74 82 25	29.411	900	81 00 00	30.000	935	87 42 25	30.578
866	74 99 56	29.428	901	81 18 01	30.017	936	87 60 96	30.594
867	75 16 89	29.445	902	81 36 04	30.033	937	87 79 69	30.610
868	75 34 24	29.462	903	81 54 09	30.050	938	87 98 44	30.627
869	75 51 61	29.479	904	81 72 16	30.067	939	88 17 21	30.643
870	75 69 00	29.496	905	81 90 25	30.083	940	88 36 00	30.659
871	75 86 41	29.513	906	82 08 36	30.100	941	88 54 81	30.676
872	76 03 84	29.530	907	82 26 49	30.116	942	88 73 64	30.692
873	76 21 29	29.547	908	82 44 64	30.133	943	88 92 49	30.708
874	76 38 76	29.563	909	82 62 81	30.150	944	89 11 36	30.725
875	76 56 25	29.580	910	82 81 00	30.166	945	89 30 25	30.741

SQUARES AND SQUARE ROOTS—Continued

No.	Square	Square Root	No.	Square	Square Root
946	89 49 16	30.757	976	95 25 76	31.241
947	89 68 09	30.773	977	95 45 29	31.257
948	89 87 04	30.790	978	95 64 84	31.273
949	90 06 01	30.806	979	95 84 41	31.289
950	90 25 00	30.822	980	96 04 00	31.305
951	90 44 01	30.838	981	96 23 61	31.321
952	90 63 04	30.854	982	96 43 24	31.337
953	90 82 09	30.871	983	96 62 89	31.353
954	91 01 16	30.887	984	96 82 56	31.369
955	91 20 25	30.903	985	97 02 25	31.385
956	91 39 36	30.919	986	97 21 96	31.401
957	91 58 49	30.935	987	97 41 69	31.417
958	91 77 64	30.952	988	97 61 44	31.432
959	91 96 81	30.968	989	97 81 21	31.448
960	92 16 00	30.984	990	98 01 00	31.464
961	92 35 21	31.000	991	98 20 81	31.480
962	92 54 44	31.016	992	98 40 64	31.496
963	92 73 69	31.032	993	98 60 49	31.512
964	92 92 96	31.048	994	98 80 36	31.528
965	93 12 25	31.064	995	99 00 25	31.544
966	93 31 56	31.081	996	99 20 16	31.559
967	93 50 89	31.097	997	99 40 09	31.575
968	93 70 24	31.113	998	99 60 04	31.591
969	93 89 61	31.129	999	99 80 01	31.607
970	94 09 00	31.145	1000	100 00 00	31.623
971	94 28 41	31.161			
972	94 47 84	31.177			
973	94 67 29	31.193			
974	94 86 76	31.209			
975	95 06 25	31.225			

INDEX

- Accomplishment quotient (A.Q.), 193.
- Accuracy, standard of, 270.
- Aims of education, statement of and adherence to, 10; Bobbitt's statement of, 10; general, 11; value of general limited, 11; must be definite, 13.
- American people, a practical folk, 6; have faith in schools, 52; support schools lavishly, 52.
- Arithmetic mean, 279; weighted, 281; computation of by short method, 283.
- "At age" and normal child, 90.
- Averages, 279.
- Ayres, criticizes Binet tests, 84-85, 93-94; handwriting scale, 181; spelling scale, 255; measurement retardation, acceleration, and elimination, 254-257; computation coefficient of correlation, 337-344.
- Bibliography. *See* end of each chapter.
- Binet, conception of intelligence, 64; age-grade method of measuring intelligence, 75; difference between individuals one of kind, 82; his later conception, 83.
- Binet-Simon tests, 67-68; synopsis of, 68-70; innovations, 72; brought constituent functions of intelligence into play, 72; kinds of mental functions brought into play, 74; provisional scale, 1905, 75; problems to be solved, 75-76; Ayres criticizes, 84-85; age at which tests should be assigned, 86-87; problems in scoring, 87-88; all-or-none method, 88; point scale for measuring, 88; number of tests to be passed each age-level, 88; with what tests shall examination begin, 89; give more than composite picture, 90; criticisms of, 93; do not determine limits of trait, 94-95; some favorable criticism of, 99; revisions and extension of, 104-122; Stanford revision, 104-107.
- Black, W. W., 36.
- Bobbitt, J. F., 10.
- Bowley, on correlation, 325.
- Bridges, point scale, 111-117.
- Burgess, measurement silent reading, 172.
- Bureau of Standards, Washington, D. C., 5, 40.
- Burt, conception of general intelligence, 65; measurement of backward children, 113.
- Business, compared with schools, 42; variables in, 42.
- Cattell, establishes psychological laboratory, 60.
- Central tendency, measurement of, Ch. X, 278-303.
- Class intervals, 274.
- Classification of pupils by educational tests, 190-192.
- Confidence of public, must cultivate, 49.
- Correlation, measurement of, Ch. XII, 324-360; need for, 324; illustrating computation of, data ungrouped, 328; data grouped and complex, 331-337; perfect positive, 331; computing by Ayres' short method, 337-344;

- graphic method of representation, 344-350; reliability of, 360.
- Creeds, educational, 12-13; scientist and, 13.
- Culture, emphasizes things of mind, 25; development of, 26.
- Curve plotting, 273.
- Danger to be avoided, 12.
- Data, on school-room problems lacking, 25; rules for tabulating, 272.
- Davenport, on correlation, 325.
- Dearborn group intelligence tests, 132.
- Defective child, compared with normal, 100-102.
- De Sanctis tests for mental defec-tion, 121.
- Deviation, computation of mean, 308; with grouped data, 310; general formula for computing, 313; standard computation of, 315; mean and standard com-pared, 317-321.
- Diagnostic tests. *See* Tests.
- Distribution, of subnormal indi-viduals, 78; table of, 275; analysis of, 275.
- Downey, on will profile, 142.
- Ebbinghaus' conception of intel-ligence, 64.
- Economy, affected through small savings, 16; applied to indi-viduals, 18; in education, 18-19; through measurements, Ch. II, 10-55.
- Educative situation, 8.
- Educational age, 192; compared to mental age, 192.
- Educational measurements, has scoffers and zealots, 43; can-not plot curve of genius, 43.
- Educational quotient (E.Q.), 192, 194.
- Efficiency, demand for, 1; in the Gary schools, 2; teachers', 2; books on, 2-3; how mechanic determines, 3; of school system, 3; American citizen, 3-4; meas-ured in higher education 6-7; working hypothesis for acquir-ing, 8.
- Energy must be conserved, 15.
- Equation of straight line of re-gression, 351.
- Error, question of, 262; compen-sating *vs.* cumulative, 270.
- Examinations, time consumed in giving, 147; attitude of teachers and pupils toward, 148.
- Experience, profiting by that of others, 39; in business, 39-40; in medicine, 40; in manufac-turing stoves, 41.
- Facts, inability to show works hardships, 32; business man and, 33; "show up" school by, 34.
- Factual basis for education, 20.
- Fechner, 57.
- Fields of educational tests and measurements, 53-54.
- Form board for measuring intelli-gence, 108.
- Freeman handwriting scale, for-mation of, 182.
- Genius, differs from intelligence, 63.
- Goddard, 100.
- Gray, handwriting score card, 182-184.
- Gregory, study of reading vocabu-laries, 237; scoring American histories, 240-247.
- Gregory-Spencer geography tests, 214; divisions of, 216; selec-tion of cities in, 217; advan-tages of tests thus designed, 221; determination of scores in, 222; objections, 224; effects of in-correct statements, 225.
- Group intelligence tests, 122-136; principles involved, 123; re-quirements for, 124; Terman, 125; National, 127; Haggerty, 129; Otis, 131; Dearborn, 132; number given, 135.

- Habermann, definition of normal individual, 78.
- Haggerty, intelligence examinations, 129; summarizes work of measurements, 140; on factors which condition success, 140.
- Hall, establishes psychological laboratory, 60.
- Hardwick, point scale, 111-117.
- Healy, picture completion test, 108; criticizes mental testing, 143.
- Human energy must be conserved, 15.
- Hume's description of metaphysical sciences, 20-21.
- Inductive method, not used in pedagogy, 24.
- Intelligence: measurement of, Chs. III-IV, 56-144; some major problems of, 56; what it is, 63-65; definitions of, 63; differs from memory and genius, 63-67; Zeihen attempts to measure, 65; as general faculty of the mind, 65; Burt's investigations of, 65; general faculty of, 65-67; inability to define does not prohibit measurements, 67; maturity of, 70; depends on correlation and interfunctioning, 73; differences in degree and kind, 82; choosing tests to measure, 83; tests must not be influenced by, 84; tests must have symptomatic value, 85; tests must not depend on use of language, 85-86; group tests of, 122-136; levels for various occupations, 132; summary and evaluation, 136-145; methods crude, 137; symposium on, 138; suggestions to teachers, 144.
- Intelligence quotient (I.Q.), 81, 192.
- Jaedorholm, experiments on sub-normal children, 83.
- James, 38.
- Johnson, on grading high-school students, 163.
- Jones, studies on spelling vocabularies, 236.
- Kant, on psychology as a science, 59.
- Kelly, F. J., on teachers' marks, 164.
- Keeping records of methods tried, 49.
- Knollin, tests "hoboes," 132.
- Knowledge, attribute of man of science, 9.
- Kuhlmann, on mental maturity, 92.
- Limited quantities make measurements necessary, 14.
- Lowell, on simplicity of tests, 124.
- Marking system, now in vogue, 150; inadequacy of, 156; inefficient, 158.
- McCall, location of zero point, 199; definition of median, 289.
- Mean, 279; computation of mean deviation, 308.
- Measurements, in physical sciences, 5; teachers should use, 9; suggestions to teachers for, 144; of school achievements new, 152; more exact make education a science, 164; by opinion, comparison, and standardized tests, 189; in other fields, Ch. VIII, 232-258; of materials of instruction 233-249; of spelling vocabulary, 233; of physical growth of children, 249-251; of school buildings, 251-253; of retardation, acceleration, and elimination, 253-258; educational compared with other fields, 264; of central tendency, Ch. X, 273-303; of variability, 304-323.
- Measures, distribution of about central tendency, 262; undistributed, 271.

- Median, 287; definitions of, 287-289; formula for finding, 290; computation of in simple distribution, 291; in complex distribution, 293; compared with mean and mode, 302.
- Medical criterion for separating subnormal from normal, 79-80.
- Mental age, coefficient of, 90-91; how to find, 106.
- Mental maturity, age of, 91-93; Terman's view of, 92.
- Meumann, conception of intelligence, 64; criticizes idea of general factor in intelligence, 66; proposed reorganization of Binet scale, 117.
- Mode, 286, 302.
- National intelligence tests, 127.
- Need for definite measurements, Ch. V, 147-179.
- Normal probability curve, 77.
- Normal frequency curve. *See* Normal probability curve.
- Normal individual, Habermann's definition of, 78; criteria for separating from subnormal, 79.
- Norsworthy, experiments on feeble-minded children, 83.
- Objections to educational tests, 43-48.
- Obstacles to rational educational reform, 28.
- Opinions: kinds and uses, 21; when worthless, 22; philosophical *vs.* knowable facts, 24-25; when no excuse for, 26.
- Otis group intelligence tests, 131.
- Paterson, scale of performance tests, 109-110.
- Pearson, experiments on subnormal children, 83; on correlation, 325-326; equation for regression lines, 353.
- Pedagogical criterion for separating subnormal from normal, 79.
- Performance tests, 86.
- Percentiles, 303.
- Picture completion tests, 107-108.
- Pintner, scale of performance tests, 109-110.
- Point scale for measuring mental ability, 111-117; distribution of tests, 114-115.
- Porteus, on mental maturity, 92; on limitations of Binet tests, 97.
- Pragmatism, 6.
- Probable error (P.E.), 306.
- Progress conditioned by ability to measure, 4.
- Psychiatrist, work of, 61-63.
- Psychological criteria for separating subnormal from normal, 79-80.
- Psychological measurements retarded, 59; made slow progress from Aristotle, 61.
- Public asking for a ledger account, 31; interested in education, 51.
- Pyle, criticizes Binet tests, 97.
- Quadrants, 349.
- Quantities measured indirectly, 266.
- Quartile deviation, 306.
- Quartiles and percentiles, 303.
- Regression lines, equation for, 351; Pearson's equation for, 353.
- Rice, J. M., 28.
- Rossolimo, mental measurements, 120.
- Rugg, definition of median, 287.
- Salt Lake City survey, 30.
- Scale of performance tests, 109.
- Scale, definition of, 159; ideal must have, 161; has been subjective, 167; some characteristics of ideal, 196; zero point of, 196; making steps equal, 202-211; must measure desired product, 211.
- Scores, value to be assigned, 226; weighting, 228; accumulation

- and difficulty, 229; interval, 289.
- Scoring, tests and treatment of measures, Ch. VII, 227; teachers' judgment in, 228; American histories, 240-249.
- School, purpose of, 18; is a state monopoly, 32; compared to business, 42; variables in, 42.
- Seashore criticizes Binet tests, 96-97.
- Secrist, definition of median, 288.
- Series, discrete and continuous, 271.
- Single variable, law of, 172; illustration of, 172-174.
- Social economic criteria for mental deficiency, 78-79.
- Spearman, conception of general intelligence, 65-66; mental maturity, 92; method of rank correlation, 362.
- Specifications, manufacturer has, 41; schools should have, 42.
- Spelling vocabulary, measurement of, 233; of Ayres' scale, 234.
- Standards, establishment of, 34-39; bricklayer has, 36; three kinds needed, 36; quantity, 36; time, 38; quality, 39; changing, 39.
- Stanford Revision Binet Scale, 104-107.
- State examination questions examined, 216.
- Statistics, general statement of, Ch. IX, 260-277; uses in other fields, 261; definition of, 267; laws of statistical regularity, 268; methods of, 269; limitations of, 270; understanding statistical formulas, 276.
- Stern, definition of intelligence, 63; coefficient of mental age, 90.
- Strayer and Engelhardt, score card, 251-253.
- Superintendents' meeting at Indianapolis, 28.
- Supervision, improved by measurements, 165.
- Talent, differs from intelligence, 63.
- Teachers, influence measured, 7; must know why changes are made, 49-50; must take initiative, 153.
- Technical scientist and educational creeds, 12.
- Terman, use of intelligence quotient, 91; on maturity of intelligence, 92; revision of Binet scale, 104-107; group tests, 125; symposium on intelligence, 138; T scale, 201.
- Tests, educational curiosities, 9; limitations of, 91; use of intelligence, 132; army mental, 132; purposes of not understood, 154; what they measure, 155; do not indicate cause of conditions, 168; how differ from other examinations, 169; how made, 169-172; when should be given, 174; number of times, 176; how improve instruction, 177; kinds most important, 178; classifying and designing, 180-185; diagnostic *vs.* general, 180; degree to which diagnostic, 181; used in Cleveland survey, 185; formal and reasoning, 185; rate and development, 186; quantity, difficulty, and time, 187; subject-matter for, 194; determined by information desired, 196; simple in application, 211; not require too much time, 212; scoring and treatment of measures, Ch. VII, 214-230; problems in scoring, 214; making a geography test, 214; values to be assigned to scores, 226.
- Testing, problems of, 100.
- Thinking, quantitatively, 5-6.
- Thorndike, 48; on general faculty of intelligence, 66; established zero in handwriting, 198; definition of median, 288.
- Time, relation to school products, 28.
- Titchener, 58; establishes psychological laboratory, 60.
- Tradition, strength of, 9.
- T scale, 199-202.

- Variability, measurement of, Ch. XI, 304-323; how measured, 305; measures of absolute, 305; coefficient of, 321; Pearson formula for, 322; Thorndike formula for, 323.
- Vocabularies, 233; reading and spelling of Oregon school children, 337.
- Wallin, criticizes Binet scale, 118.
- Waste, teachers should eliminate, 9; elimination of, 14; not peculiar to American schools, 14; opportunities for in education, 15; how business man eliminates, 16; much and varied in education, 17.
- Watt, 4.
- Weber, 58; experimental work in psychology, 58-59; laws, 58.
- Woody, arithmetic scale, 195.
- Wooters, 21.
- Wundt, 57; establishes psychological laboratory, 57-58; effects of laboratories, 60.
- Yerkes, point scale, 111-117.
- Zeilen, attempts to measure intelligence, 65.
- Zero, point of scale, 161, 196; in composition, 197; in penmanship, 198; McCall establishes, 199; in T scale, 200.

To avoid fine, this book should be returned on
or before the date last stamped below

10M-6-40

JUL 1945

FEB 20 1943

MAR 31 1943

~~MAR 31 1943~~

JUN 1943

DEC 1945

MAR 1 1946

JUL 19 1946

JUL 19 1946

350

371.27 .G822 C.1
Fundamentals of educational me
Stanford University Libraries



3 6105 033 373 270

Basement

371.27

G82

271

267

180

360989

OF EDUCATION

